ENSURING THE CONTEXT VALIDITY OF ENGLISH READING TESTS FOR ACADEMIC PURPOSES (EAP) IN OMAN

A.A.S. AL ISMAILI

PhD

2015

Ensuring the Context Validity of English Reading Tests for Academic Purposes (EAP) in Oman

Anwar Amur Salim AL ISMAILI

Submitted for the Degree of

Doctor of Philosophy

Faculty of Social Sciences

University of Bradford

2015

Abstract

Anwar Amur Salim Al Ismaili

Ensuring the Context Validity of English Reading Tests for Academic Purposes (EAP) in Oman

Keywords:
Assessment, Context Validity, Second Language (L2), Reading, Validation, English for Academic Purposes (EAP), Cognitive Processes, Task Setting, Careful Reading, Expeditious Reading

Students entering academic programmes are frequently overwhelmed by the demand for extensive reading and comprehension of information derived from multiple and contrasting sources. This entails both careful and expeditious reading. The latter has been generally neglected in research and has not been the focus of many tests. Both types of reading were investigated in this study through a validation process of the summative English reading test for academic purposes taken at the end of the Foundation Programme in Oman. In particular, context validity was established through focusing on the interaction between the linguistic demands and task setting parameters and also the cognitive processes through which the students engaged with the test tasks.

To establish the context validity of the test, this study adopted Khalifa and Weir's (2009) model which not only embraced the complex and multi-componential nature of reading but also provided a workable validation framework.

A multi-strategy approach was adopted. A natural experiment utilising Verbal Protocol Analysis captured the cognitive processes through which students engaged in reading. Automated analysis software and opinions of expert judges were used to compare test passages with text extracts drawn from first year academic courses. Correlation tests and factor analysis revealed these cognitive processes and established the robustness of the Khalifa and Weir (2009) model, which was thus validated in a second language context. Passages in the foundation tests were found to be generally representative of academic texts although certain features such as abstractness were under-represented.

Acknowledgments

This thesis has been a journey. All journeys have their moments of exhilaration and inspiration. These are often rare experiences when compared with the many moments of perplexity and desert experience where the way ahead is often obscure. I have discovered that it is through persistence in these moments that insight and inspiration are born.

In the first place, I wish to put on record how trust in Allah has sustained me throughout the long journey and has given me the inner strength to persevere in order to complete my thesis.

I have to make a special mention of the generosity of the government of the Sultanate of Oman, represented by the Ministry of Manpower, in providing me with this once in a lifetime opportunity for personal and professional development. I hope that my final thesis goes some way to repaying that kindness through making a contribution to the current practice of assessment in Oman.

Next, I acknowledge the invaluable support I received from Dr. Tim Green who was so patient and was always there to keep me on track to reach the various milestones of the journey. In particular, I want acknowledge his willingness to go the extra mile with me. Likewise, I would like to express my gratitude to Mercedes Webb who always managed to keep me grounded by considering the practicality of different aspects of the journey.

A special word of gratitude is due to Professor Tony Green who guided me through the desert regions of statistical analysis to reach the verdant pastures of research in second language testing. From him I have learned much about the need for precision in my thinking and writing which have always challenged me to go yet further along the road.

Many other people, too numerous to be singled out, have supported me on this journey. I am grateful to my family for their understanding and support

and especially to my late grandmother for her constant support and prayers and also other family members who always remembered me in their thoughts and prayers.

Raya, my wife showed great patience and understanding and never once complained about the long hours of research often lasting deep into the night. My two daughters must also have wondered what I was doing locked for hours in the seclusion of my study.

Another person I wish to single out is John Kinsella my proofreader. In truth he was a fellow traveller on the journey as he gave generously of his time and support. I hope he has enjoyed the journey just as much as I have.

Last but not least, I wish to express my gratitude to the students and expert judges who participated in this research and gave freely and generously of their time. I trust that they now appreciate the fruit of their efforts.

Dedications

To the memory of my dear grandmother Thuraya, my mother Raya and my father Amur, who despite their own lack of educational opportunities, nevertheless, in their own ways, inspired me to take my first steps in learning and the pursuit of knowledge

May Allah be merciful to them and rest their soul in peace

Table of Contents

List of Figures

Chapter 6

Chapter 8

List of Tables

Chapter 1

Chapter 2

Chapter 3

Chapter 4

Chapter 6

Chapter 7

Glossary

ACCESS for ELLs……	Assessing Comprehension and Communication in English State-to-State for English Language Learners
AWL……..	Academic Word List
CAE……..	Certificate in Advanced English
CAEL……...	Canadian Academic English Language assessment
CEFR Level A…….	Basic user
CEFR Level B…….	Independent user
CEFR Level C…….	Proficient user
CEFR……	Common European Framework of Reference for Languages
CET……..	College English Test
CPE……..	Certificate of Proficiency in English
CT……..	Colleges of Technology
EAP……	English for academic purposes
EFL……	English as a Foreign Language
ESL……	English as a Second Language
ESOL…….	English for Speakers of Other Languages
ESP……	English for Specific Purposes

ETS……	Educational Testing Service
FCE…….	First Certificate in English
FP……....	Foundation Program
FYA……	First Year Academic
GEPT…….	General English Proficiency Test
GFP……..	General Foundation Programme
HCT……...	Higher College of Technology
HEI…….	Higher Education Institution
iBT…….	Internet-based TOEFL
ICT…….	Ibra College of Technology
ICT……	Information and Communication Technology
IELTS……	International English Language Testing System
IT……	Information Technology
KET……	Key English Test
L1……..	First Language
L2……...	Second Language
LEE……...	Level Exit Exam
MBA……...	Master of Business Administration
MCQ……..	Multiple Choice Questions
MELAB……..	The Michigan English

	Language Assessment Battery
PET……	Preliminary English Test
PGCE……	The Postgraduate Certificate in Education
RQ…….	Research Question
RSQ……..	Research Sub-Question
SAQ……..	Short answer questions
SPSS…….	Statistical Package for Social Sciences
TEEP…….	Test of English for Educational Purposes
TOEFL…….....	Test of English as a Foreign Language
TTR……	The Type/Token Ratio
VPA……	Verbal Protocol Analysis
WPM……	Words Per Minute

Chapter 1 Introduction

1.1 Background of the research

As English has become the lingua franca of communication in commerce, trade, engineering and in many other fields, it has become common practice in many countries to deliver college or university courses through the medium of English, even when English is not the spoken language of the people. In Oman, students undertake a Foundation Program (FP) in the Colleges of Technology (CTs) in preparation for progression to degree or higher diploma courses delivered through the medium of English. Entry to degree or higher diploma studies is dependent, inter alia, on reaching a satisfactory level in the final test of proficiency in English for Academic Purposes (EAP) on the FP. The Omani government, in numerous public policy statements, recognized and stressed the important and fundamental role of the English language worldwide as the language of science and technology and as an effective tool for modernization embracing sociolinguistic, socioeconomic, socio-cultural, historical and political factors (Al-Issa, 2002). This view was echoed later by Al-Husseini (2006) who also stressed the importance of the English language for national development within a global economy.

Oman is an Arab Gulf country located on the Arabian Peninsula. The strategic Strait of Hormuz, which controls access to the important international trade route of the Arabian Gulf, is located in Oman. The country has an oil rentier economy but in recent decades the government has seen the need for diversification to minimize dependency on oil (Muscat Media Group, 2014). Oman has a population of approximately 3.3 m (CIA Office of Public Affairs, 2015), but one third of this population consists of expatriates. This is a relatively small population, which is mainly located in the capital Muscat and a small number of cities. Thus, many parts of Oman are sparsely populated. The median age of the population in 2014 was 24.9 years (Office of Public Affairs, 2015), which compares with a median age of 40.0 years for

the UK population in the same year (Office for National Statistics, 2015). This means that half the population in Oman (approximately 1.1 m) is below the age of 25 and, as a consequence, approximately 40,000 students graduate from secondary schools each year.

The main driving force behind the imperative for the inclusion of English in the foundation programs is the process of ‘Omanisation’ (Al-Issa, 2005, 2006) which sees English as a fundamental tool for its success. This implies a policy which aims at the gradual replacement of expatriate skilled workers by nationals (Al-Issa, 2006) whereby indigenous students would acquire the necessary knowledge and skills to be able to fulfill key roles in the national economy.

The Omani economy began to modernise following the accession of His Majesty Sultan Qaboos to the throne in 1970. This was characterised by investment in the development of infrastructure such as roads, electricity and housing. Priority was also given to the development of a health service and the modernisation of the existing education system (IHS Global Insight Inc., 2012, 2015).

Later modernisation has taken on a more global dimension. As is the case with many countries, Oman would have to develop itself as a world class economy in order to have a sustainable economy capable of competing in global markets. Recent investment opportunities have included multi-billion dollar investments for continuing to upgrade infrastructure, the development of high-speed rail links and aviation expansion, the development of Information and Communication Technology (ICT) projects, tourism and international hotels (The Public Authority for Investment Promotion and Export Development, 2013). Of particular importance is the multi-billion oil refinery development at Duqm on the Southern coast, a strategically located development for the secure supply of oil worldwide from the Gulf region which continues to be characterised by political instability (Special Economic Zone Authority Duqm, 2013).

Towards the end of the 1980s, modernisation became increasingly concerned with the funding of human resource development, whereby Omani people would be able to advance to third level qualifications. This included some scholarships for study abroad. Table 1.1 shows how scholarships for degrees and qualifications have been steadily increasing over recent years (Higher Education Admissions Centre, 2015):

Table 1.1 Numbers of enrolled students in the HEIs according to HEI/scholarship and academic year[1]

HEI/Scholarship	2009/2010	2010/2011	2011/2012
Sultan Qaboos University	2717	2747	3107
Colleges of Applied Sciences	1929	1997	2100
Technical Colleges	6225	8222	10625
Institute of Sharia Sciences	171	187	139
Institute of Health	595	637	605
Internal Scholarships & Grants	2922	3087	9978
External Scholarships & Grants	194	229	1444
Total	14753	17106	27998

World-class economies are essentially knowledge-based economies and this is reflected in the later stage of the modernisation of Oman by the expansion of education and training and the acquisition of skills, including English that are essential for such a development. This has resulted in the expansion of

[1] This table includes only secondary school graduates and does not include other scholarships granted by various providers such as ministries and private sector.

the Colleges of Technology (CTs) over recent decades. The CTs were founded in 1984 and within a decade had grown to 5 colleges. More recently two further colleges were added. This expansion is evidence of the growing number of students aspiring to study in the CTs. In fact, over the past ten years, all sectors of adult education have seen growth, but none more than the CTs, where student intake has more than trebled in the last ten years. This fact, coupled with the government's strategy of Omanisation to create a knowledge-based economy by 2040, has resulted in an explosion of young people in further and higher education. Evidence of the phenomenal growth of the CTs is seen in the fact that in the academic year 1993-1994 there were only 674 students in the CTs. This number had risen to 2000 in the 2000-2001 year and to 31,463 by 2012-2013. Of these approximately 11,500 were on the FPs (see Table 1.2).

Table 1.2 Number of students at Colleges of Technology in semester 1 2012/2013

Programme	Male	Female	Total
1 Post Foundation	11215	8191	19406
2 Foundation Programme	8571	3208	11779
On job Training	177	0	177
Total	19963	11399	
Grand Total	31362 students in 7 CTs		

Adopted from (Directorate General for Technological Education, 2012)

CTs have become the largest higher education provider in terms of student capacity. For example, in 2012/2013 approximately 28,000 students progressed to further or higher education and, of these, 37.4% (10750) enrolled in the CTs (Higher Education Admissions Centre, 2015), making the CTs by far the most popular destination of post-secondary students. The vast

majority of these students are placed each year on the Foundation Programme (FP) prior to commencing their more academic or technological education. This is because in their secondary education, English was taught as a discrete subject and other subjects were taught through the medium of Arabic. Cooper (1984) has commented on the advantage enjoyed by secondary students whose subjects were delivered through English over students as in the case of this study whose secondary education was through their first language. Thus, many of the Omani students are likely to be impeded in advancing to study through English at academic level. Cooper has pointed towards the particular difficulties encountered by students who have not been taught through the medium of English at secondary level such as their lexical deficit among other factors.

Thus, these students need to become proficient in English before advancing to their chosen courses as these are delivered through the medium of English (Ministry of Manpower, 2004). The necessary skills are developed through English for Academic Purposes (EAP) provision integrated into the FP. The FP also develops students' academic skills such as functional mathematics and IT.

Thus, the success of the FP is of critical importance for the economy as it impacts on the progression of some 12,000 foundation students annually from these 7 CTs. Also, the success of the Omanisation programme is ultimately dependent on the success of these students, which, in turn, hinges on the validity of the English language tests.

This study focuses on the Level Exit Exam (LEE - the final summative assessment) of English proficiency at the end of the FP, the scores of which are used for making decisions concerning individual students' abilities to undertake academic or technological studies through the medium of English. This LEE tests all the main skill areas of English including speaking, listening, writing and reading, each being weighted of equal importance (25%) in the scoring. The pass mark is based on the overall score.

1.2 Statement of the research problem

The fact that the LEE in the CTs in Oman is such a high-stakes test implies that it needs a consistent validation process. A validation process provides empirical evidence that an individual test task measured what that task was intended to measure and that the inferences made on the basis of the test scores are valid (e.g. Bachman 1990; Weir, 2005; Green, 2014). For example, a particular test task may have been designed to measure grammatical resources for comprehension but, if test takers were no less likely to get the correct answer by another means, then the validity of that test item would be in doubt. Consequently, it could not be conclusively inferred that test takers were able to use grammatical resources as an aid to comprehension. Since this would be required for reading at academic level, establishing validity is important for many reasons including the avoidance of possible negative impacts (e.g. O'Sullivan, 2012). There is currently no such validation process in existence in the CTs in Oman. The need for empirically-based validation studies is well supported in the literature which emphasises the importance of such studies being evidence-based (e.g. Messick, 1989; Kane, 2002; Hughes, 2003; Bachman, 2004; Weir, 2005; Davies, 2011).

Many international and local second language tests conduct regular validation studies which adduce evidence for stakeholders of the test takers' proficiency in the second language (L2) for the purpose for which the test was designed. Some examples of international tests which conduct validation studies include International English Language Testing System (IELTS) (IELTS, 2015), Test of English as a Foreign Language (TOEFL) and Internet-based TOEFL (iBT) (ETS, 2015) and the Michigan English Language Assessment Battery (MELAB) (CaMLA, 2015). In fact, some of these tests have been the subject of a large number of studies of various kinds, for example TOEFL which is "…a heavily researched test, with roughly 100 research and technical reports on earlier versions of the TOEFL and a monograph series, starting in 1997, that reported on developments, inter alia, in the TOEFL 2000 projects" (Alderson, 2009, p.626). What is true of international tests is also true of many, though by no means all, localised

tests which are now applying good practice of quality of second language tests and validation, for example, the Canadian Academic English Language assessment (CAEL) (The CAEL Assessment Office, 2015), the General English Proficiency Test (GEPT) in Taiwan (The LTTC, 2015) and ACCESS for ELLs in the US (WIDA Consortium, 2015).

Nevertheless, despite these exemplars of good practice, many local tests do not have a rigorous system of validation. The need for such a system of evidence-based validation studies has been officially recognized for second language testing in Oman: “The HEI must demonstrate that the chosen assessment method is effective in determining whether the student has attained the required learning outcomes” (OAAA, 2008, p.8). Despite such a mandate from the Oman Academic Accreditation Authority for foundation programs, there is a paucity of research into second language assessment in general in the CTs in Oman. In fact, to date, only two scholarly studies have focused on the College of Technology in Oman: Al-Hinai (2011) and Al-Husseini (2004). Neither of these studies focused on second language assessment, despite its pivotal importance for the successful implementation of Omanisation and also bearing in mind that the CTs have been running since 1984. Thus, the need for such second language validation studies in the CTs in Oman is a compelling one.

Whilst there is an urgent need for validation studies into all skill areas of English language testing, this study focuses on reading and comprehension. Once students progress to academic or technological studies, they are quickly confronted with much reading from many sources. Coping with multiple sources of information places great demand on skills of reading and comprehension. However, it is not always possible to read many of the sources in great detail. Often, what is required is the ability to read expeditiously so that the essence of a particular passage or chapter can be grasped in a relatively short time. English language tests have tended to focus on careful reading (slow and deliberate where sometimes the meaning is incremental) to the neglect of expeditious reading (fast and efficient skimming, scanning and searching) (Weir, 2013). In fact, attention has been

drawn to the neglect of expeditious reading in second language testing, e.g. in IELTS, where there has been a predominant emphasis on careful reading despite evidence that expeditious reading skills and strategies were also of critical importance, not only for L2 university students but often also for L1 students, many of whom encountered problems with applying these skills effectively (Weir et al., 2009). However, this was based on studies in a first language (L1) context where English was the everyday spoken language, albeit with second language (L2) students. The current study addresses this research gap for L2 students in a L2 context by a more balanced approach that gives prominence to expeditious reading alongside careful reading. Furthermore, previous studies (Weir et al., 2009; Green et al., 2010) have been focused on students beginning more academic types of subjects (e.g. sociology, law etc.), whereas this study is focused on technological subjects (e.g. engineering, information technology (IT)…etc.) being taken at academic levels. Although English is still important, it is worth noting that subjects such as engineering and IT may be less dependent on proficiency in English than would be the case with more discursive subjects such as sociology or law, where the ability to engage in argumentation and the interpretation of arguments would be important. Additionally, this study pays more attention to the processes employed by the test takers in deciding which type of reading is appropriate in a given context. This concurs with recent theories of reading and assessment which have focused on the cognitive skills which are involved and the strategic choices which readers make about the type of reading which is appropriate in a given situation (e.g. Cohen, 1994; Weir, 2005; Grabe, 2009; Moore et al., 2011). Assessments of both types of reading are mandated in the standards of General Foundation Programme (GFP) in Oman:

- “Read a one to two page text and identify the main idea(s) and extract specific information in a given period of time.
- Read an extensive text broadly relevant to the student’s area of study (minimum three pages) and respond to questions that require analytical skills, e.g. prediction, deduction, inference” (OAAA, 2008, p.10).

The expected outcomes of the assessment clearly intend that both careful and expeditious reading should be tested along with the decisions as to which type of reading would be appropriate for given passages.

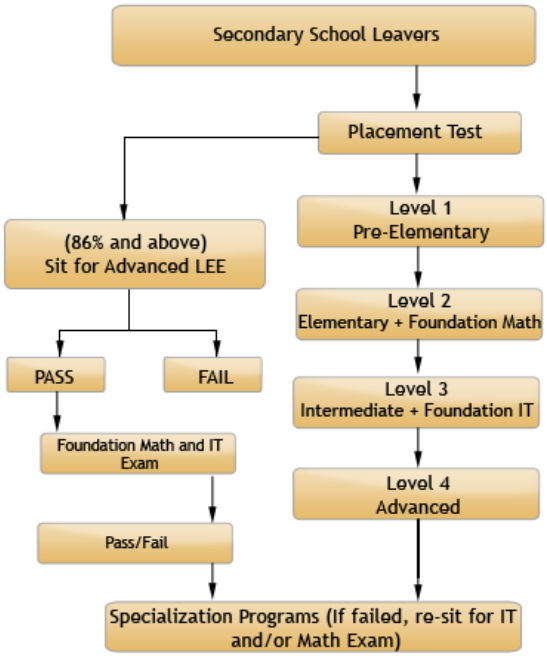
In addressing the research gaps, there is a number of existing models of test validation that serve as frameworks for empirical investigation. Out of these models, there emerges a conceptualisation of reading as being multi-componential in nature. In the literature review, different models are discussed and critiqued and the case is made for adopting the Khalifa and Weir (2009) model. This model is founded on Weir's (2005) validation framework in which different aspects of validity (e.g. cognitive, context) were identified and which also took into account the interaction between contextual parameters (linguistic demands and task setting) and cognitive processes (test taker) for interpreting abilities underlying test scores. Empirical studies have confirmed the importance of both of these aspects in L2 tasks, e.g. Tavakoli's (2009) significant finding, based on three studies of task difficulty, in which both teachers and learners had identified cognitive and linguistic demands as the two most important factors relating to task difficulty and, based on these findings, concluded that educators needed to take both factors into account in programme design and testing.

Many important subsequent empirical studies have been founded on Khalifa and Weir (2009), for example Weir et al. (2009), Green et al. (2010), Wu (2011), Weir et al. (2012) and Yanagawa (2012). In fact, this model is unique in combining a theory of language with a theory of validation (Kantarcioğlu, 2012). The theory of language embraces the different types and modes of reading; the theory of validation considers the various parameters for measuring different aspects of validity. However, the authors have drawn attention to the need for more studies to be conducted into the context validity of reading tests due to the gap in the current research (Weir et al., 2009). Accordingly, the focus of this research is on the context validity of the reading test. Additionally, this research addresses another research gap by investigating a number of task settings which have not been included in previous investigations of context validity.

1.3 Purpose of the research

The purpose of this research is to investigate the context validity of the reading section of the Level 4 Exit Exam (LEE) of English in the CTs in Oman. Level 4 represents the highest level in the FP. There are four levels in the FP and students are placed at the most suitable level following a placement test. Each level has a final summative test (LEE) and the test at the end of the fourth level is the test on which entry to the academic courses is determined. The process is presented in Figure 1.1:

Figure 1.1 Foundation Programme (FP) structure in the Colleges of Technology (CTs) in Oman



In particular, this study investigates the validity of the reading section of this high-stakes test in terms of its context validity following Khalifa and Weir's (2009) model. Thus, judgements based on test scores using this validation research are considered as providing more robust evidence of the test taker's ability to successfully progress towards academic reading and study. Simply stated, this necessitates an evaluation of how representative the reading test tasks in LEE are of the texts and tasks involved in reading and comprehension at academic level.

1.4 Aims of the research

The current research has the following principal aims:

- To identify the cognitive processes by which students engage with the texts and tasks in the reading tests.
- To examine how closely the reading texts in LEE tests reflect those encountered at first year academic (FYA) level.
- To investigate how closely the reading tasks in LEE tests reflect those encountered at first year academic (FYA) level texts.

1.5 Overview of the thesis

The structure of the current research is presented in 8 chapters as outlined below. *Chapter 1* presents the research problem set against the background of the study in the CTs in Oman. It also introduces the underlying model which frames the current research and its purpose and aims.

In *Chapter 2* the different theories of reading are surveyed in search of an appropriate model for investigating the testing of reading in the English for Academic Purposes (EAP) test in Oman. The nature of reading is explored and the variables by which the contextual features may be assessed are identified. Based on this, various gaps in the research are identified.

Chapter 3 presents the methodology and research design which are employed to address the main research question in order to achieve the aims and objectives of the study. The main research question and research sub-questions (RSQ) emerged from the literature review. Automated text analysis was used to determine how representative the LEE tasks were of texts typically encountered at First Year Academic (FYA) level. For variables not directly amenable to this method, the decisions of a panel of expert judges were taken into account. Additionally, test takers took part in a natural experiment using Verbal Protocol Analysis (VPA) in order to identify the various cognitive processes involved in their reading and comprehension. The findings from both methods were expected to provide complementary evidence for answering the research questions. Issues of validity and

reliability are discussed, ethical considerations are addressed and limitations of the methods are acknowledged.

In *Chapter 4* the findings of the natural experiment are presented and the cognitive processes through which the test takers engaged with reading are identified.

This is followed in *Chapter 5* by a critical discussion of the main findings with reference to the relevant literature.

In *Chapter 6*, the data from the automated text analysis software and expert judges were analysed from the perspective of answering the two research sub-questions which were concerned with determining how representative the reading texts and tasks were of the texts typically encountered at first year academic level.

In *Chapter 7*, the findings from the previous chapter are critically discussed in the light of the relevant literature. Each of the contextual features is discussed in turn and a picture of the overall findings is presented.

The final chapter, *Chapter 8*, draws conclusions following the two discussion chapters. The key findings are brought together and convergences and divergences are noted and discussed. The complex and multi-componential nature of reading is confirmed and a case is made for the adoption of the Khalifa and Weir (2009) model which was validated here in a L2 context. Implications for theory, test design and for curriculum and programme planning are stated and recommendations are made. Limitations of the research are also identified.

Chapter 2 Literature review

2.1 Introduction

The focus of this study is on reading in English for Academic Purposes (EAP) in Omani higher education. This literature review considers different theories and models of reading in order to identify a model that is appropriate for the evaluation of tests of EAP reading in Oman. Such a model should prove useful as a basis for evaluating the reading level four test (LEE) tasks in the Foundation Program (FP) and comparing these with the actual texts that students would be likely to encounter when studying through the medium of English on their chosen academic programmes in their first year at college (FYA) (see Chapter 1 Section 1.3).

The review establishes which reading processes are important in the context of EAP reading and so can be used as a basis for assessing how well the test tasks measure these processes[2].

The review is divided into eight sections. Following this introductory section, the nature of reading is explored from different theoretical perspectives. In the third section, the principal models of reading are presented and critically evaluated. This is followed, in the fourth section, by a consideration of issues involved in validity and validation. The fifth section considers Weir's (2005) test evaluation framework and presents an argument for its inclusion as the theoretical model underlying the current research. The sixth section focuses on context validity which is explored in this research and paves the way for a presentation, in the penultimate section, of the contextual features and variables which affect the nature of reading. The final section summarises

[2] Attention has been drawn by Purpura (1999, p.7) to the "considerable confusion" resulting from attempts to differentiate between strategies and processes in some studies. Operationally, he views strategies as "conscious or unconscious techniques, behaviours or activities that an individual invokes in language learning, use or testing" whereas processes are viewed as referring to "individual latent variables underlying strategy use during language learning, use or testing" (Purpura, 1999, pp.7–8). Purpura's distinction between strategy and process is followed throughout this study except when directly citing an author who uses an alternative understanding.

the chapter and sets the scene for the research design and methodology which is presented in the following chapter.

2.2 Nature of reading

Navigating into the history of testing of reading in the 20th century, Weir (2013) presents contrasting perceptions about how reading is treated in the most influential English speaking and testing countries, UK and USA. Weir (2013, pp.110-111) discusses how the prominence of reading as a separate skill has been slow to develop in the UK in the 20th century. Teaching and assessment in the UK have focused on:

- Decoding at clause and sentence level
- Lexis
- Attention to careful local reading at clause and sentence level
- Research on comprehension (scant until 1950s)
- Vocabulary as a major determinant for difficulty of understanding reading and intelligence testing.

A major gap identified, however, is the absence of focus on expeditious reading. Weir acknowledges, "…expeditious reading (fast, efficient, selective reading) has not been explicitly tested in many high stakes examinations in the United Kingdom to this date, despite its inclusion in CEFR reading descriptors for academic purposes at the C1 levels" (Weir, 2013, p.112).

Although the earlier period focused on reading for L1 learners, the importance of reading in English for L2 and for academic purposes was highlighted by a number of writers (e.g. Carrell et al., 1988; Enright et al., 2000). That reading is a highly complex activity has been emphasised by Alderson (2005), who then asserts that "…the complexity of reading in a foreign language is equally well established" (Alderson, 2005, p.119).

Reading is not directly observable. In fact, there is considerable debate about the very nature of reading (Hubley, 2012). Goodman (1988) viewed it as 'a receptive language process.' For him it is a psycholinguistic process in that it "…starts with a linguistic surface representation encoded by a writer

and ends with meaning which the reader constructs" (Goodman, 1988, p.12). Goodman actually views reading as a cyclical rather than a linear process within which "the readers leap toward meaning" and "The cycles are telescoped by the readers if they can get to meaning" (Goodman, 1988, p.15). Thus, Goodman is being more descriptive than prescriptive, the point being to explain the processes a reader could possibly go through to arrive at meaning but always bearing in mind that meaning is the object of the process and that it can be grasped in such a way that it is sometimes not necessary to go through each cycle. Thus, Goodman's model is comprehensive in that it represents the processes that are involved for a variety of different kinds of readers and purposes. However, his model, although pointing to the complex nature of reading, is still not sophisticated enough to embrace the numerous decisions and activities involved in reading that are seen in later models. For example, it does not consider the strategic judgment the reader makes regarding the type of reading, either careful or expeditious, which is such an important feature of later models such as Weir's (2005) model. Nor is it a framework that can be easily applied in research in second language testing. Rather, it is a more theoretical base on which teaching and testing of reading plans can be established (Goodman, 1988).

A distinction has been made between different types of reading. One such distinction concerns 'active and interactive' reading (e.g. Carrell, 1988; Samuel and Kamil, 1988); Anderson and Pearson, 1988). Although Carrell (1988) considers the reader's background knowledge in active reading, the focus is primarily on decoding the text in order to comprehend, as closely as possible, the meaning inherent in the text. Interactive reading is a more dynamic approach to the meaning of the text whereby there is an interaction between the reader and their processes and characteristics, e.g. background knowledge, and the text itself. Moore et al. (2011) see the activity of reading as being a happy medium between two opposite points of a continuum, one of which is reader-oriented and the other is more text-oriented. The reader-oriented end of the continuum places more emphasis on what the reader brings to the text by way of comprehension, whereas at the other end of the

continuum, the meaning to be comprehended is seen as lying largely in the text itself and is there to be discovered by the reader. They cite Carroll (1964) as an example of the text oriented perspective where a test task is seen as an attempt at measuring the extent to which the test taker has been able to define the pre-existent meaning that resides in a printed text. The alternative view is that texts do not have a single definitive meaning but a range of possible meanings. This is especially the view in some branches of the humanities. Yet, Moore et al. (2011) state that there is a reluctance, even in a postmodernist context, to fully accept the consequences of this position which is to deny that there is any objective account of the meaning of the text but only the interpretation that the reader brings to the text. This has led to the interactive view that both processes are part of reading (e.g. Bernhardt, 1991; Weir et al., 2009). Although in the Omani context most test takers are progressing to more exact sciences such as engineering, IT and mathematics (see Chapter 1 Section 1.2) where texts contain less ambiguity than more interpretive sciences such as sociology and literature, it is still important in the context of EAP that students are able engage with test passages in an interactive way. Consideration needs to be given, in test design for reading in the Foundation Programme, to the prior knowledge and experience that these test takers can be reasonably expected to have. This includes what they have studied at earlier stages (for example, levels 1 to 3 in the FP, see Chapter 1 Section 1.3). However, the challenge facing test designers in Oman is to construct test tasks that challenge the test takers to draw on their prior knowledge in order to complete the test tasks.

Another issue relates to whether reading is a unitary skill or a group of empirically distinguishable skills (Adam, 1990; Alderson, 2000). Throughout the literature, it is evident that there was an emphasis on discrete items in language testing and assessment in general and this also applied to testing reading. However, some authors (e.g. Harris, 1968: Brindley, 2001) have been critical of this approach to testing and have highlighted its inadequacies. It was evident that there was a need to shift to a more integrative approach as argued by authors such as (Groot, 1975).

Alderson (2000) emphasises the socio-cultural dimensions of reading, first

language aspects and reading and cognition. Olson and Torrance (2009) also highlight the socio-cultural differences involved for learners for whom the language is, not only not their first language, but is also significantly different in structure such as learning English by people whose first language is, for example, Arabic or Mandarin where a totally new orthography is involved. Grabe (1991) disagrees with the often cited difficulty that orthographic differences between a student's L1 and English would present certain difficulties. In fact, while it may be true for beginners, there is less evidence that this is true for advanced readers of English as "…it appears that direction-of-reading differences cause little difficulty… Punctuation and spacing of written forms also differ across languages but do not seem to disadvantage any linguistic group" (Grabe, 1991, p.387).

Since reading is not directly observable, it is, therefore, difficult to define. Essentially, reading involves the perceiving and understanding of written language. As early as 1961, Lado offered a functional definition of reading as "…grasping meaning in the language through its written representation" (Lado, 1961, P.223). The emphasis here is on the language itself and on the graphic symbolization of that language (Lado, 1961). Lado's idea of "grasping meaning" remains a central issue for language acquisition. However, since Lado's contribution, other important factors have come to be recognised: "Reading calls for the reader to actively supply meaning to text on a continual basis" (Cohen, 1994, p.212). There is a shift towards a consideration of the cognitive processes which are involved for individual readers in terms of schemata formation in relation to the test taker's background knowledge and content knowledge. Additionally, there are test takers' characteristics, which have an influence on the comprehension of what is read. "Age, gender, socio-cultural background, learning background inter alia may all affect how a test-taker reacts to a particular test" (Brown and McNamara, 1992, p.68). The importance of test taker characteristics is also supported in (e.g. Bachman, 1990; Bachman and Palmer 1996; Saville, 2000; O'Sullivan, 2012). In practical terms, obtaining evidence for test development and validation of test takers' characteristics is also considered of crucial importance for Cambridge ESOL (Khalifa, 2005) and has clear

implications for test design of EAP in the FP in Oman. Following Weir (2005), this implies that test design in Oman needs to take fully into account the physical/physiological, psychological and experiential characteristics of the test takers. This includes an understanding of the world of young Omani adults (aged 18 to 21), both male and female, coming from diverse backgrounds ranging from urban to remote rural areas.

Alderson and Urquhart (1984) assert that there is no conclusive answer to the question “What is reading?”; rather than answers, more questions are posed. Perhaps the activity we call reading, which implies comprehension, is best described rather than strictly defined. The type of description that answers a question such as “what happens when I read?” or “what are the processes that occur when reading takes place?” might lead to more descriptive definitions. Even until recent times in the literature of reading, Grabe (2009) acknowledges the complexity of defining reading as actually reflected in the nature of reading: “…it is evident that no single statement is going to capture the complexity of reading” (Grabe, 2009, p.14). A solution, he suggests, is to direct attention to look into fluent readers and the characteristics of their reading, followed by considering the processes of reading and their complex combination (Grabe, 2009). One may pose the question: why focus on fluent readers? Day and Bamford explain “What is true for fluent readers—that slowing down and paying conscious attention to recognizing words interfere with the construction of meaning—is even more true for beginning readers. The disruption is such for beginners that the link between decoding process and the comprehension processes may be severed” (Day and Bamford, 1998, p.15). Accordingly, as in the case of test takers’ characteristics considered above, it would also be advantageous for test designers in Oman to investigate fully the processes through which test takers engage with their reading test tasks.

In view of the foregoing discussion which has served to highlight the complexity of reading, it follows that assessment of reading will also be challenging (Cumming, 2008). The students in the Foundation Programme in the Colleges of Technology are certainly confronted by a challenge indeed, especially as they are second language readers in a second language

context (see Chapter 1 Section 1.2). These students are generally weaker as the more able students coming from secondary school have already been selected for overseas scholarships or local universities. They are generally operating at the level of decoding but at a very local level. This raises the question of how realistic the current educational targets are, given that the expected benchmarks are hard to attain in the time scale. However inadequately prepared these students are at entry to the main programme, by the end of the first year, many of them are coping adequately, so it could be the case that, in order to make progress in their chosen vocational area, they become much more motivated to work at comprehension and meaning in reading.

2.3 Models of reading

Prior to considering the assessment of reading, it is first necessary to survey the different theories relating to the nature of reading itself. A useful outline is presented in (Athey, 1971; Carrell, 1988; Samuels and Kamil, 1988; Grabe, 1991; Clapham, 1996; Hudson, 2007; Grabe, 2009). It is most helpful to present these in a tabular form below:

Before 1970:

Model
▪ L2 reading viewed as an adjunct to oral skills (Carrell, 1988) ▪ Reading as decoding sound-symbol relationships and mastering oral dialogues (Carrell, 1988, p.2) ▪ Recognition of the importance of background knowledge, particularly socio-cultural meaning in L2 reading comprehension. (Carrell, 1988, p.2)
▪ Audio-lingual theories were common in the 60s ▪ Psycholinguistic theories which involve: Selective process Using knowledge Prediction

In the 70s:

Model
Smith's (1971) model cited in Samuel and Kamil (1988, p.24): ▪ Reading is now emphasised as distinct from other elements of language. ▪ Reading as hypothesis driven ▪ It works despite the redundancy inherent in language (Samuel and Kamil, 1988, p.24).
LaBerge and Samuel's (1974) model: ▪ Emphasizing automaticity of component process ▪ Linear bottom-up, focuses on rapid processing and word recognition ▪ How the reader reads rather than comprehension (Hudson, 2007, pp.36-37).
Rumelhart's (1977) model: ▪ Interactive model

- "Emphasising flexible processing and multiple information sources, depending upon contextual circumstances" (Samuel and Kamil, 1988, p.24)

Clarke and Silberstein (1977) cited in Carrell (1988, p.57):

- Reading is an active process of comprehension also mentioned in Carrell (1988, p.3)
- Teaching involves strategies
- Guessing defines expectations
- Inferring
- skim ahead
- Pre-reading activities (vocabulary and grammar)

A model of ESL reader by Coady (1979):

- Extended the theory to L2 learners
- Three components:
- process strategies
- background knowledge
- conceptual abilities
- Beginners use process, more advanced use more abstract
- confirming and predicting seen as important

The 1970s reading focused on reading to another:

- L2 reader active in processing information while sampling text (Carrell, 1988, p.3)
- Growing dissatisfaction with audio-lingual method (Carrell, 1988, p.3).

Kintsch and van Dijk (1978) model cited in Samuel and Kamil (1988, p.24)

- Emphasising comprehension to the exclusion of word identification (most other models, Including Rumelhart's, seem to have a bias for explaining word identification." (Samuels and Kamil, 1988, p.24)
- Linear information processing models, but later interactive models developed (Samuel and Kamil, 1988, p.25).

Widdowson (1978, 1983) cited in Carrell et al. (1988, p.3) Emphasised communicative aspects of ESL “Viewing second language reading as an active process in which the second language reader is an active information processor who predicts while sampling only parts of the actual text” (Carrell et al., 1988, p.3).

The 1980s focused on English as a second language (ESL):

Model
Stanovich (1980) model: “A key concept is that “a process at any level can compensate for deficiencies at any other level” (Stanovich, 1980 cited in Samuel and Kamil, 1988, p.32) [such compensation occurs at either level deficiency see Stanovich 1980, p.36] - an interesting twist to the Rumelhart model” (Samuel and Kamil, 1988, p.24) Stanovich’s contribution: “…under certain conditions poor readers exhibit greater sensitivity to contextual constraints than do good readers. They do in circumstances where featural, orthographic, and/or lexical knowledge…sources are weak in comparison to syntactic and semantic knowledge. The reason good readers are sometimes less sensitive to contextual effects is that their knowledge sources for these lower-level processors are seldom weak…” (Samuel and Kamil, 1988, pp.32-33).
Just and Carpenter (1980) cited in Samuel and Kamil (1988, p.24): “…account for comprehension processes based upon studies of eye movements.” (Samuel and Kamil, 1988, p.24)
Goodman’s model (1988): Reading as an active process. “Whether or not one agrees with Goodman that there is indeed a *single* reading process, or that miscue analysis is the best or even an

appropriate way to access this process, Goodman's model sets the stage for approaching reading an *active process*" (Carrell et al, 1988, p.9)
Samuels and Kamil's (1988) model: ▪ top-down and bottom-up processing ▪ interactive ▪ interacting hierarchical stages, rather than discrete linear stages (Carrell et al., 1988, p.9)
Anderson and Pearson (1988) cited in Carrell et al. (1988, p.10): Interactive but includes schema theory "…comprehension involves the interaction between old and new information" (Carrell et al., 1988, p.10).
Grabe's model (1988): ▪ Relating models developed for native to the domain second language reading (Carrell et al., 1988, p.10) - interactive ▪ Focuses on "… the relation of the reader to the text, or it may focus on the processing among the various component skills and stages, or it may even focus on features of the text itself" (Carrell et al., 1988, p.10)

From the tables above, it is worthwhile to point to the critique of Rumelhart (1977) where linear models are adjudged to be inadequate for accounting for "…a number of occurrences known to take place while reading" (Samuel and Kamil, 1988, p.27). These occurrences include the following observations:

1. Letter recognition is significantly faster if the letters form a word rather than a non-word
2. When a word recognition error is made, the substituted word tends to be a similar part of speech
3. Semantic knowledge influences word perception
4. Syntax depends on the context of the word
5. Interpretation depends on the context of a text segment (Samuel and Kamil, 1988).

Linear models tend to neglect any consideration of a feedback loop (Samuel and Kamil, 1988), often described as bottom-up and are rather mechanistic in approach. The main point of critique of linear models is that they fail to account for processes known to occur in reading which suggest that a much more interactive process is involved: "…the lack of the feedback loops in the early bottom-up models, [made it] difficult to account for sentence-context effects and the role of prior knowledge of text topic" (Samuel and Kamil, 1988, p.31). This disadvantage of bottom-up approaches is also emphasized in (Grabe, 1988). Underlying the bottom-up approaches are the assumptions of behaviourist psychology that somehow the brain is stimulated by words on the page. In contrast, the top-down models generally tend to embrace cognitive theory, especially with guessing and predicting (Samuel and Kamil, 1988). Nevertheless, there is a growing consensus that reading does not come down on the side of either choice, but rather that it is a complex activity involving both top-down and bottom-up processes.

The models of reading are derived from the taxonomies on which decisions are made regarding the various skills and sub-skills that constitute reading. Munby's (1978) taxonomy is seminal in this respect. He identified 266 skills which were grouped into 54 categories, making it a rather complex scheme. Due to its complexity, it has been critiqued in reviews by (McKay, 1980; Wilkins, 1980; Davies, 1981; Mead, 1982) on the grounds of its impracticality and the dubious nature of the model from a theoretical perspective. More recently, Grabe and Stoller (2002) reduced the number of skills to more manageable proportions. They proposed that reading can be usefully described as:

- Searching for factual points
- Skimming
- Knowledge seeking
- Relating what is read to other knowledge,
- Relating reading to writing
- Reading for the purposes of criticism
- Reading for general understanding

This is in contrast to Enright et al.'s (2000) attempts to simplify it to just four elements:

- Reading for information,
- Reading to comprehend at basic level,
- Reading for learning,
- Reading for integrating information from a number of texts.

Grabe (2009) discusses 11 distinct models, which he categorises into four types:

- Descriptive models
- Models that collect assessment evidence
- Experimental and behavioural models
- Models where performance is based on statistics

Included are many of the models presented in the table above.

Effective models of reading, according to Goodman (1988, pp.20-21), should account for the following factors:

Firstly, characteristics of:

- Text, text structure, text length
- The reader
- Syntax and grammar
- Semantic systems
- Memory
- Perception
- Orthography

Secondly, conditions of reading:

- purpose,
- task setting,
- third party influences,
- context

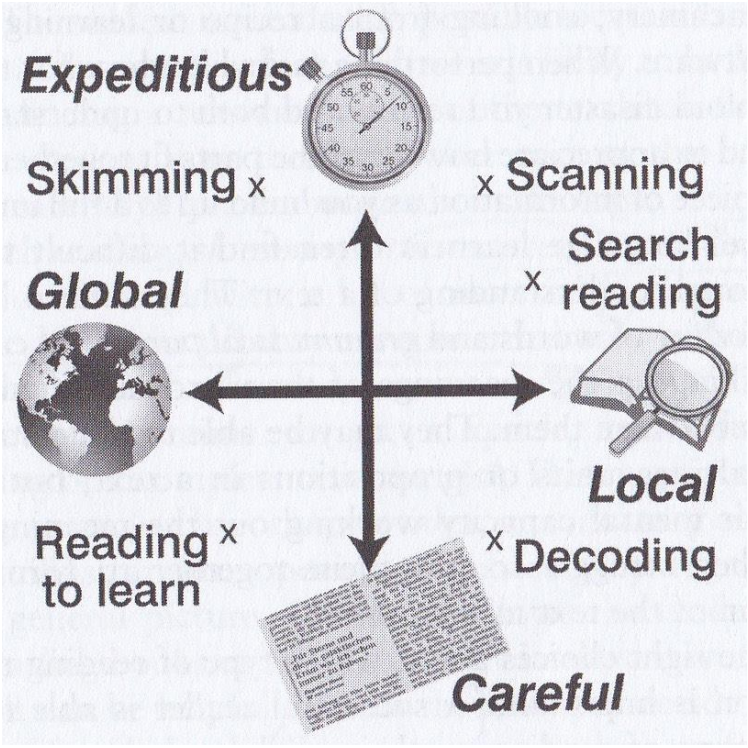
Many of these features are also found in Weir's (2005) test validation framework, a development of the earlier model of Urquhart and Weir (1998). This was further developed into the socio-cognitive model of Khalifa and

Weir (2009) which examined reading and its underlying processes. A socio-cognitive model as devised by Khalifa and Weir (2009) is one in which "…the abilities to be tested are demonstrated by the *mental* processing of the candidate (the cognitive dimension)" and also one which equally embraces "the use of language in performing tasks as a social rather than a purely linguistic phenomenon" (Khalifa and Weir, 2009, p.4). Thus, this model represents a unified approach in ensuring the validity of a test. However, within the remit of this study, the focus is on context validity (see Chapter 1 Section 1.2)

As has been pointed out in the introductory chapter (see Chapter 1 Section 1.2), a special feature of Khalifa and Weir's (2009) model is the way in which it has brought together a theory of test validation and a theory of language. The genesis of Khalifa and Weir's (2009) model is now discussed.

An important development in models of reading had already appeared in Urquhart and Weir (1998) who distinguished between two types of reading: careful and expeditious, a distinction which was also utilized by Moore et al. (2011) in their study of IELTS academic reading. This is illustrated in Figure (2.1):

Figure 2.1 Types of reading



Adopted from Urquhart and Weir (1998) cited in Green (2014, p.99)

The model contains a double axis between expeditions and careful reading and the global-local dimension. This distinction between expeditious and careful reading is more important in the context of this study, as students at third level need to be able to make reference to a vast amount of literature and, consequently, must be able to utilise processes of expeditious reading as well as being able to read important passages or chapters or even a larger section more carefully: " Careful reading is intended to extract meaning from presented material at a local or a global level, i.e. within or beyond the sentences right up to the level of the complete text or texts. The approach to reading is based on slow, careful, linear, incremental reading for comprehension" (Weir, 2013, p.113). Careful reading, therefore, involves decoding at the local level and reading to learn at the global level.

Expeditious reading, however, is time constrained and the aim is skimming to get the gist of a text at the global level or scanning for specific information at the local level. In essence, expeditious reading is characterised by fast

selective and efficient reading carried out with the objective of locating required information either by capturing the main ideas of the whole passage or, more simply, by pinpointing specific information. In contrast to the slow, linear and incremental nature of careful reading, expeditious reading may involve skipping part of the text in order to focus on what is perceived to be relevant to the reading purpose. However, recent research using eye-tracking as a method for identifying cognitive processes in reading (Bax, 2013), concluded that the most successful students made use of expeditious reading.

Identified and stressed as one of the processes which fluent readers employ, rapid (expeditious) reading is defined "…in the sense that we read most materials at about 250-300 wpm" (Grabe, 2009, p.14). Building on many studies, Weir (2013) underlines that "…for many readers reading quickly, selectively and efficiently posed greater problems than reading carefully and efficiently" (Weir, 2013, p.113). Continuing, he cautions that "Expeditious reading of continuous prose to access desired information in a text is difficult because it demands rapid recognition which is contingent upon sufficient practice in reading in the target language" (Weir, 2013, p.113). The difficulty of which Weir speaks in this extract was borne out by Canadian-based research involving L2 graduate students who were taking a MBA programme (Raymond and Parks, 2002). Most of the students reported that they had read 50% or less of their MBA textbooks claiming that they had no time to read them carefully and slowly. However, self-reporting revealed that a small number of students in this study had developed coping strategies to deal with the heavy load of reading. These strategies did not involve a decision to either read carefully or expeditiously. Rather, they decided to discard whole sections of the programme based on the judgment that they were already sufficiently familiar with the subject on the basis of their prior career development. Nevertheless, this is still not the optimal solution, as their assumption that their prior career experience was such that they could ignore whole sections of a programme at master level might well have proven to be a fallacious one. Similar findings were reported in Spack (1997). If this was problematic for L2 students at postgraduate level in an English speaking

context, an a fortiori argument points towards the added difficulty facing the technology students in Oman in a second language context. All of these students studied English as a discrete subject at secondary school level (see Chapter 1 Section 1.1). More is required of them as they move to third level education where subjects are delivered through the medium of English. Cooper (1984) drew attention to the difficulty facing third level students who are required to study through the medium of English without having done so at secondary level. The challenges such students face is due largely to their lexical deficit and, in the case of students whose first language is Arabic as in this study, there is the added difficulty of the opaqueness of sound-symbol correspondences in English compared to Arabic (Saigh and Schmitt, 2012). Thus, Omani students are challenged by expeditious reading which requires rapid word recognition. This implies that Omani foundation students need to practice reading in a time-constrained context in order to develop rapid word recognition. The importance of even a 15-minute practice session every week could improve speed, confidence and comprehension (Chang and College, 2010).

Accordingly, in this research, importance is attached to skills of expeditious reading of Omani test takers in view of the extensive reading which will be required of them once they progress to FYA.

Part of the problem referred to in the previous paragraph can be related to the effect that slow word recognition can have on comprehension (Laberge & Samuels, 1974; Perfetti, 1985; Adams, 1990, 2004; Samuels, 2004). Each comments on how slow word recognition affects both the accuracy and completeness of comprehension. Moreover, in the context of this study, slow word recognition would render expeditious reading very difficult, if not impossible. Cushing-Weigle and Jensen (1996) pointed out that, even after taking a course designed to improve skills of reading, L2 students at an American university were still reading at a very slow rate (100 w.p.m).

Urquhart and Weir's (1998) model was further developed by Khalifa and Weir (2009) and presented in matrix form as in Table (2.1):

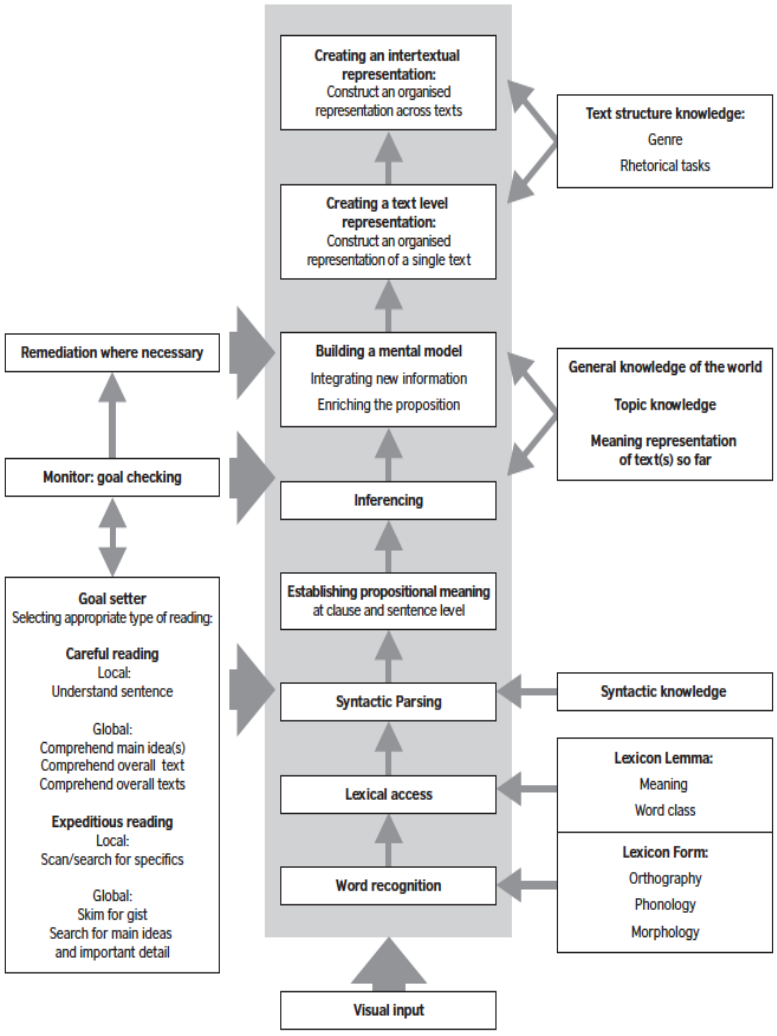
Table 2.1 Types of reading of Khalifa and Weir's (2009) model

	Global level	**Local level**
Careful Reading	▪ Establishing accurate comprehension of explicitly stated main ideas across sentences ▪ Making propositional inferences ▪ Establishing how ideas and details relate to each other in a whole text ▪ Establishing how ideas and details relate to each other across texts	▪ Establishing accurate comprehension of explicitly stated main idea or supporting details within a sentence ▪ Identifying lexis ▪ Understanding syntax
Expeditious Reading	▪ Skimming quickly to establish: discourse topic and main ideas, or structure of text, or relevance to needs ▪ Search reading to locate quickly and understand information relevant to predetermined needs	▪ Scanning to locate specific points of information

Adopted from Weir et al. (2009, p.101)

The matrix is based on the vertical axis divided between careful and expeditious reading and the horizontal axis divided between global and local levels. The underlying assumption for this model is a multi-componential understanding of reading and the assessment of reading (Weir et al., 2009). The advantages of such an understanding of reading has been pointed out by Moore et al. (2011) as being a more dynamic model with the capacity for generating a range of reading modes. Weir et al. (2009) make the distinction between global and local levels based on the need for comprehending information within the sentence and beyond the sentence. Alderson (2000) supports this view. This is also reported in many EFL studies, for example Gordon's reading test by Israeli students, (Gordon, 1987, cited in Cohen, 2012, p.98). This model was articulated in Khalifa and Weir (2009) shown in Figure (2.2):

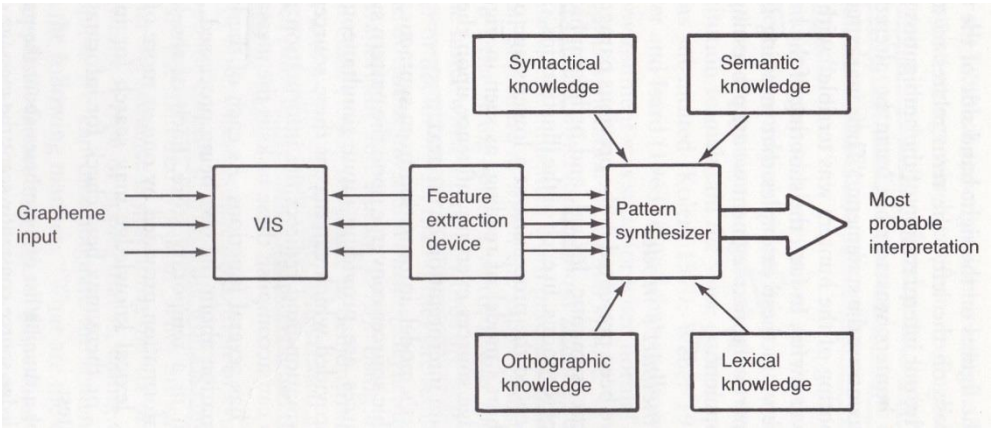
Figure 2.2 Khalifa and Weir's (2009) model of reading



Adopted from Weir et al. (2009, p.102)

In the column on the left, it is noted that there is the addition of a third activity, namely “goal setter” alongside careful and expeditious reading. Goal setter corresponds to selecting the appropriate type of reading in a goal-oriented setting. An earlier model, Rumlehart (1977), cited in Anderson and Pearson (1988) also shows reading as multi-componential and is presented in Figure (2.3):

Figure 2.3 Rumelhart's (1977) model



Adopted from Samuel and Kamil (1988, p.30)

What is missing in Rumelhart's (1977) model, however, is the important distinction between global and local levels that has helped in the categorization of the skills as in Khalifa and Weir's (2009) model. Additionally, the processes involved in 'synthesizer' are not fully delineated and also the notion of goal setter has not been included. However, Rumelhart's model seems to be mechanistic and linear. It almost equates the cognitive processes of reading and comprehension to that of a micro-processor reading an "input string" (Samuel and Kamil, 1988, p.29) in a process of hypothesis testing which continues until the most likely hypothesis is confirmed. Goodman (1967) refers to this as "the psycholinguistic guessing game" (Goodman, 1967, p.126). Nevertheless, there are elements of an interactive approach in his model by the inclusion of the four types of knowledge: syntactical, semantic, orthographic and lexical. However, the functions of these four are reduced simply to hypothesis checking. They appear as merely *attendant* skills to the 'pattern synthesizer'. Moreover, the nomenclature implies that the cognitive process is being modeled on the micro-processing unit of a computer. However, the later models, especially Khalifa and Weir (2009), include these four components as *central* to the 'levels of processing' (the central column) but include other processes (Weir

et al., 2009). That process does indeed begin with visual input but, the cognitive processes that take place are much more complex than is suggested by Rumelhart's model. The four linguistic components are included in addition to a number of other important processes. This works in line with how we read: "When we read, we coordinate rapid and automatic word recognition, syntactic parsing, meaning formation, text comprehension building, inferencing, critical evaluation, and linkages to prior knowledge sources. We do this seemingly without effort and with all processes synchronizing in time" (Grabe, 2009, p.14).

An important aspect of Khalifa and Weir's (2009) model is the strategic decision that the reader must make between expeditious and careful reading which they call a 'goal-setter'. It is clear, however, that this is not a simple choice. In fact, it is a highly developed metacognitive strategy (Purpura, 1999; Bachman and Palmer, 2010), which is more appropriately designated as 'discernment' in this study, as the term 'goal setter' does not adequately convey the processes involved in this metacognitive strategic decision. Accordingly, the notion of discernment is now explored and defined and some practical examples of its use are presented.

With reference to Green (2014), where metacognitive skills are seen as including the 'choice' the reader makes regarding expeditious reading or more considered careful reading, it should be pointed out that students in academic settings do not usually exercise this choice as a reflex action. These settings according to Grabe "require us to synthesize, interpret, evaluate, and selectively use information from texts. Moreover, we often encounter competing or contradictory information on a regular basis. It is a fact of modern life that almost any issue or topic can be discussed, addressed or argued from multiple viewpoints and it is routinely our task to decide among these alternative sources of information. How we learn to negotiate this world of print and achieve our goals is a large part of many professional and academic lives" (Grabe, 2009, p.5). Thus, it is expected that most students gradually develop the ability to select appropriately one or other of the permutations between careful and expeditious reading at either global or local level. However, it is a learning process and, at times, students

may choose careful reading where expeditious reading would be more appropriate. This is quite different from a test task where a student realises that a certain amount of work in reading and comprehension is required in a time-bound situation. In reading a text in the academic situation, there is still the constraint of time but not in the sense that is imposed in the examination situation. While a student may quickly work out, in the context of an examination, which permutation to use, in the course of textbook reading, no such clues are presented. Learning to choose wisely the type of reading for a particular university text involves an informed judgment, which is more aptly described as 'discernment'.

The Oxford English Dictionary defines discernment as "…to come or to know or recognize mentally esp. something that is hidden or obscure <the inductive apprehension of a truth imperfectly>" (Murray et al., 1933, p.644). It is this notion of 'imperfectly apprehending' that seems to describe the experience of the reader encountering an academic text and first, through skimming and scanning, comprehending it imperfectly leading to a strategic decision as to whether such an imperfect comprehension is appropriate in this situation or whether the text now needs more careful reading to comprehend it in greater depth. The etymology of the term discernment traces its roots back to Medieval Latin where it had the meaning of "*dis* = apart + *cernere* = to sift" (Murray et al., 1933, p.644). Splitting and sifting is exactly what is going on with attempts to comprehend most academic texts; metacognitive processes which imply also "a self-management or executive capacity" (Purpura, 1999, p.6). As in most management scenarios, situations encountered are not always exactly predictable and learning takes place as a result of errors as much as from successes. The text, in itself, does not provide this sort of self-management capacity but rather, this is a skill which grows and develops especially in early academic studies. Thus, it is much more in line with a reader-oriented ability similar to those presented in the recent models of reading (as discussed above in Section 2.2 and Section 2.3). It should also be noted that efforts to simulate real life scenarios in a test situation are not always as effective or authentic as designers would wish. Faced with exercising discernment in a given test task or question is

quite different from discerning the way ahead in reading an academic text where the student must rely on their own initiative. Actually, there are many situations where students do not exercise discernment in test situations, as the way the test is set up and the knowledge of the time to be allocated to a task already give the students the clue as to which type of reading to execute.

The informed judgment required for this splitting and sifting calls for a considerable degree of reading experience of different kinds of texts for different purposes. However, the splitting eventually requires what Grabe (2009) calls "*reading to integrate*", which he considers to be a major academic purpose for reading: "The effort to build a strong organising frame in reading to learn is increased significantly when there are multiple texts that refer to related information but they may present conflicting or incompatible facts and explanations. In the case of multiple texts, the reader must decide how to create his or her own organising frame for the information because none is provided by the combined texts. In the case of a long, complex text, the information may have been presented through multiple organising frames (comparison-contrast, descriptive listing, problem-solution)" (Grabe, 2009, p.9).

Once again, reading to integrate involves more than a simple judgment; involved are the students' efforts to develop an organising frame in order to manage their own reading and comprehension and this is more appropriately conveyed by the term discernment rather than the concept of goal setter. Goal setter, as conceptualised by Khalifa and Weir (2009), is essentially a matter of selecting a mode of reading with the overall goal in mind. It does not seem to include the additional elements of self-management, splitting and sifting and integration which constitute processes of discernment.

One consideration in discernment is the nature of the text itself – can it be comprehended by getting the gist of it through skimming, scanning and summarising? Furthermore, discernment also relates closely to the purpose for which the reading is being done. If after a cursory quick reading, it is decided by the reader that the purpose would be better served by a more

careful reading of the text, then the decision must be in that direction for a more nuanced and analytical comprehension of the text. This view is well supported by Green (2014): “As well as choosing the type of reading that suits their purpose, good readers also monitor and review their comprehension. They form judgments about whether they have achieved an adequate understanding” (Green, 2014, p.101). Green’s use of the word judgment clearly implies that this is a strategy of discernment. The final monitoring and review - self-management processes - are clearly elements of discernment related to the overall purpose for which the reading is being done.

In fact, Green (2014) adds a number of other factors involved in this discernment strategy. He asserts that understanding the meaning of words and sentences is not sufficient; the reader needs to “build up a *conceptualisation* of the overall meaning of the text including grouping the relationships between the different ideas expressed” (Green, 2014, p.102). He goes on to comment on the need to be able to connect ideas of one author with those of another and comments on how challenging a skill this is due to the fact that the ideas of different authors are not often explicitly presented. Additionally, it is necessary for the reader to be able to align or compare and contrast these different views in a coherent way (Green, 2014). This echoes the earlier emphasis placed on reading to integrate and the creation of organising frames highlighted in Grabe (2009).

Grabe (2009) distinguishes good readers from poor readers by claiming that, although they use the same strategies, good readers use these strategies more effectively, and clearly this must imply processes of discernment. Some practical examples taken from effective readers are presented in the table below from Grabe (2009, p.228):

Table 2.2 Strategies used by engaged readers

#	Strategies used by engaged readers
1.	They read selectively according to goals.
2.	They read carefully in key places
3.	They reread as appropriate.
4.	They monitor their reading continuously and they are aware of whether or not they are comprehending the text.
5.	They identify important information.
6.	They try to fill in gaps in the text through inferences and prior knowledge.
7.	They make guesses about unknown words.
8.	They use text-structure information to guide understanding.
9.	They make inferences about the author, key information, and main ideas.
10.	They attempt to integrate ideas from different parts of the text.
11.	They build interpretations of the text as they read.
12.	They build main-idea summaries.
13.	They evaluate the text and the author and, as a result, form feelings about the text.
14.	They attempt to resolve difficulties.

Adopted from Grabe (2009, p.228)

It is of interest to investigate the extent to which these strategies are incorporated into the Omani Foundation level tests, particularly at level four, the highest level in the Foundation Programme, immediately prior to the commencement of academic studies.

In summary, the various models of reading and comprehension support the view that they are highly complex operations not directly amenable to observation. Although unitary and simple models have been presented, the prevailing consensus emerging from the literature about theories and models is that reading is multi-componential and interactive, especially for academic purposes, and operates at both local and global levels. Consequently, the testing of EAP reading needs to reflect the very complexity of reading and its multi-componential nature. Test tasks need to assess the test taker's ability to not only read carefully and expeditiously when appropriate, but also to recognize the difference. For example, for expeditious reading, tasks that call for skimming and scanning within a reasonable time frame are required (the issue of time and speed is discussed under Task Setting below Section 2.7.2), and for careful reading, usually longer more discursive texts should be included in order to align with the texts usually encountered at academic level or first year. Thus, it is imperative, for the purposes of this study, to assess how the different types of readings at each level are represented in the test tasks in the Omani colleges. In this regard, attention must be paid to the difficulty of Omani students using processes such as scanning. This is because their L1 is Arabic which is a highly inflected language and characterised by salience features which call for considerable attention to be given to the meaning of individual words rather than focusing on a group of words (Palmer et al., 2007). Thus, these students may find it difficult to apply skimming and searching at a global level under time constraint. Randall (2009) has drawn attention in particular to the difficulty of test takers whose L1 is a highly inflected language such as Arabic is and offers this as an explanation for why they tend to be more successful at reading expeditiously at local level (scanning) but not at global level (skimming or searching).

2.4 Validity and validation

Most authors consider validity as being concerned with the issue of whether a test actually measures what it purports to measure (e.g. McNamara, 2000; Hughes, 2003). Validation is concerned with a process of "…collecting

evidence about the validity of the inferences that may be made on the basis of the assessment results" (Green, 2014, p.242). This concurs with the development of the earlier understanding of validity whereby Messick (1989) extended validity to include interpretation and consequences of scores as evidence. However, Messick's model of validity was complex and difficult to operationalise (O'Sullivan, 2011). A major step forward was Weir's (2005) test validation framework which took into account the interpretation of test scores based on evidence. This evidence was based on the test score reflecting a certain language ability which was considered to result from the interaction between the cognitive processes by which the test taker engaged with the test task and the contextual features of the test task (Weir, 2005). In contrast with Messick's model, Weir's model was one which could be operationalized and took into account different aspects of validity such as cognitive, context and consequential aspects for providing empirical evidence (as is discussed below in Section 2.5). This research is based on the Khalifa and Weir (2009) model which was a further development of Weir (2005). However, this research focuses on context validity in particular as was stated in the research problem in the introduction chapter (see Chapter 1 Section 1.1 and Section 1.2).

2.5 Test validation frameworks

The most prominent of the various frameworks in the field of language assessment are now briefly considered. Earlier frameworks such as Bachman's (1990) and Bachman and Palmer's (1996) which although quite comprehensive, were lacking in empirical underpinning (O'Sullivan and Weir, 2011). In the specific context of testing of reading, Fulcher is adamant that: "until models are empirically investigated, there is a danger that tests designed on the basis of under-defined models may lead to construct under-representation in the test, and this represents a major threat to test validity" (Fulcher, 1998, p.282). Additionally, O'Sullivan and Weir claim that another disadvantage of Bachman's model is that it fails to address "…the treatment of the cognitive processing dimensions (as versus metacognitive) of the

various skills components…when considering its use for test development purposes, especially where attempts are made to define different levels of language proficiency" (O'Sullivan and Weir, 2011, p.15). It is of interest here to find a set of clear and unambiguously stated variables that can be used to ensure the context validity (contextual features and cognitive processes) of tests in the FP by comparing them with texts taken from first year academic in order to achieve the aims of this research by providing empirical evidence (see Chapter 1 Section 1.4). However, Alderson and Kremmel (2013) have highlighted the difficulty of arriving at such a robust set of variables for L2 reading but nevertheless underline its necessity for identifying and pinpointing learners' strengths and weaknesses. In the light of this observation and Fulcher's (1998) plea for an empirical investigation of adopted models, this research attempts to validate the Khalifa and Weir (2009) model for use with L2 students in a L2 context (Chapter 4). Khalifa and Weir (2009) is a development of the earlier framework of Weir (2005).

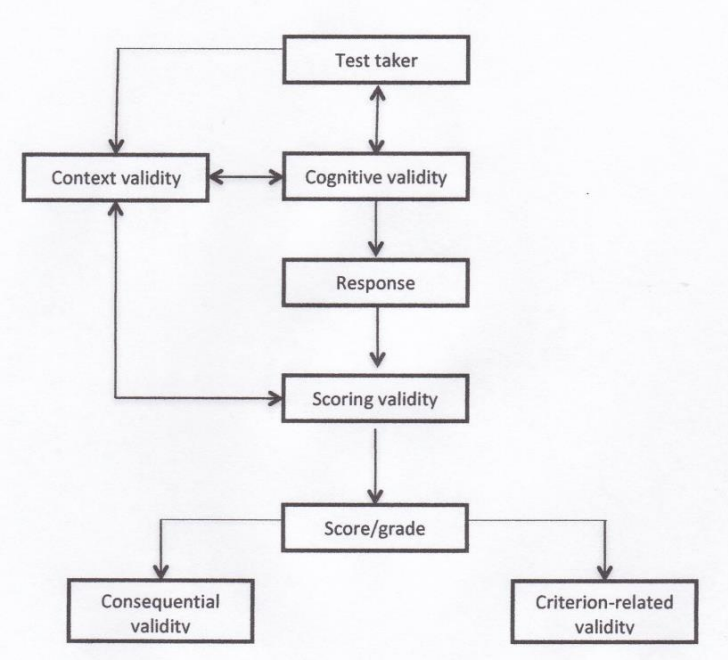
Weir's (2005) framework draws on previous models, particularly Bachman's, but also provides an empirical basis for that framework (Weir et al., 2009; Green et al., 2010). For these reasons, Weir's framework is important as an underpinning for the Khalifa and Weir (2009) model which is preferred as a basis for this study. The strengths of Khalifa and Weir's (2009) model have been pointed out by Bax: "Khalifa and Weir's model describes cognitive processing in reading in terms of different levels of complexity, with, for example, lexical processing as the least complex, and intertextual reading as the most. Khalifa and Weir's model is therefore particularly valuable in that it operationalizes the concept of cognitive processing in reading, and proposes in a way amenable to empirical investigation a hierarchy of cognitive processing complexity in reading" (Bax, 2013, p.443).

Moreover, Weir et al. (2009) have acknowledged that context validity had not been fully empirically supported and this presented a gap in research that needed to be properly addressed. A good example is seen in task setting, the importance of which is acknowledged in Weir's (2005) framework, yet had not been included in their empirical testing, a fact that was readily

admitted by Weir et al. (2009). One of the purposes of the current research is to address this gap by the inclusion of variables which touch on these parameters in view of their stated importance to the overall context validity of test scores: “no serious studies have been undertaken in which the focus is on the contextual parameters and the cognitive processing” (Weir et al., 2009, p.100). In fact, only two studies since 1995 have exclusively focused on the IELTS reading module, let alone on finer elements involved in reading assessment (Weir et al, 2009). Weir et al. (2009) make a strong plea for further research in this area. Since then, in the context of the current research, two important studies have emerged which embraced contextual parameters and cognitive processes in the testing of reading for academic purposes: namely Weir et al. (2009) and Green et al. (2010). In the same vein, Hudson argues that “The role of the reading task and text type have yet to be thoroughly researched in second language studies. More research is needed in this area” (Hudson, 2007, p.73).

It is however useful to first consider Weir’s (2005) validation framework due to its contribution to the model on which this research is based. Weir’s test validation framework (2005) (Figure 2.4) assumes that cognitive processing always takes place within a given context:

Figure 2.4 Weir’s (2005) test validation framework



Adopted from O’Sullivan and Weir (2011)

The context of the current research is the ability to read and comprehend text readings at college level. Thus, the test task performances should enable inferences to be made about performance in the real world situation of academic reading. This amounts to activities and tasks expressed in terms of cognitive processes within contextual parameters.

Situational authenticity (reflecting contextual features) and interactional authenticity (reflecting cognitive processes) in Bachman and Palmer's (1996) model were seen as of crucial importance in devising useful test tasks and these are closely related to the contextual and cognitive validities in Weir's framework (Weir et al., 2009; Khalifa and Weir, 2009). Whilst acknowledging the difficulty of embodying contextual features due to various constraints, not the least of which is time, Weir et al. (2009) strongly advocate that as many contextual features as possible should be incorporated into the test tasks. In fact, Douglas goes even further by asserting that "…it is important to be as specific as we can in establishing contexts for our test takers" (Douglas, 2009, p.24).

Weir (2005) divides these contextual features into three categories: task setting, linguistic demands, and task administration. Task setting includes purpose, response format, known criteria, weighting, order of items, and time constraints. Linguistic demands consist of linguistic variables, both input and output, and include discourse mode, channel, text length, writer-reader relationship, nature of information, content knowledge, lexical, structural and functional. Finally, administration setting includes physical conditions, uniformity of administration and security. Task administration merits a separate treatment and is therefore not included in this research.

O'Sullivan and Weir (2011) see " a "symbiotic" relationship between the context, cognitive and scoring aspects of validity, a relationship that is seen by these authors as representing a unified approach to establishing the overall construct validity of a test" (O'Sullivan and Weir, 2011, p.21). This is supported by Taylor who views the model as capable of providing us with "… a useful heuristic for reflection and action in the area of language diversity"

(Taylor, 2009, p.149). A consideration of all these aspects of validity is beyond the scope of this study which is focused on context validity. Accordingly, context validity is explored with reference to the relevant literature in the following section.

2.6 Context validity

"Context validity is concerned with the extent to which the choice of tasks in a test is representative of the larger universe of tasks of which the test is assumed to be a sample" (Weir, 2005, p.19). This view of representativeness is well-supported in the literature although it is referred to as content validity (e.g. Davies, 1990; Weir, 1990; Cohen, 1994; Alderson et al., 1995; Brindley, 2001; Hughes, 2003; Fulcher and Davidson, 2007; Akbari, 2012; Farhady, 2012; Flowerdew and Miller, 2012; Green, 2014). Weir (2005) makes the point that context validity is a more all-embracing term covering all the realities (a viewpoint which is adopted in this research and, for that reason, the term context validity is used in this sense throughout the thesis except when citing other authors directly). Weir sees context validity as accounting for "…the social dimension of language use" (Weir, 2005, p.19). This introduction of social dimension use is derived from Messick's seminal work: "To appraise how well a test does its job, we must inquire whether the potential and actual social consequences of test interpretation and use are not only supportive of the intended testing purposes, but at the same time are consistent with other social values" (Messick, 1989, p.8). This resonates with the new approach to teaching literacy which "…views reading as part of a social and cultural process that establishes, maintains, or changes social relationships, and frequently emphasizes the non-school practices of literacy" (Hudson, 2007, p.56). It leads to a more comprehensive view of reading by sharpening the focus on "…these social relations rather than reducing concepts of reading simply to the reader-text processes" (Hudson, 2007, p.56).

Fulcher (1999) sees content validity as relating "…not only to test content, but also to test format. The target domain should be described in the needs

analysis in terms of both content and the language use situations in which the student will be expected to survive, and the analysis is then translated directly into test content and task type" (Fulcher, 1999, p.222). Controversy surrounds the establishment of a representative sample, what abilities and skills should be included and to what extent these should be included. McNamara (2000) comments on the need for the test construct to relate to real world tasks. He draws a direct relationship between the content of a test and the construct from which the test content is chosen. McNamara (2000) draws specific guidelines by which test content can be established and derived from the domain of the test construct that was already established. These are:

(1) "It can be defined operationally, as a set of practical, real-world tasks...
(2) Alternatively, the domain can be defined in terms of a more abstract construct..." (McNamara, 2000, p.25).

Anastasi (1988) proposes important guidelines in order to establish content validity:

1. "the behavior domain to be tested must be systematically analysed to make certain that all major aspects are covered by the test items, and in the correct order;
2. the domain under consideration should be fully described in advance, rather than being defined after the test has been prepared;
3. content validity depends on relevance of the individual's test responses to the behavior area under consideration, rather than on the apparent relevance of item content" (Anastasi, 1988, cited in Weir, 2005, pp.19-20).

These considerations were taken into account in establishing context validity in this study by an 'a priori' analysis of the test tasks before the actual test event as advocated by Khalifa and Weir (2009) (see Chapter 6 and Chapter 7). Furthermore, Anastasi's (1988) point about taking into account not simply the content of the task but the significance of test takers' responses to the tasks was borne in mind in evaluating the processes involved in responding to the various test items (see Chapter 4 and Chapter 5).

However, both Fulcher (1999) and Young (2008), in considering validity, caution against the inclusion of irrelevant variables. Additionally, Fulcher counsels about the need to avoid 'contamination' in test specification either by reducing under-representation of tasks and items or through eliminating irrelevant variances to avoid construct-irrelevant sources of variances entering into score interpretation (Fulcher, 1999). Such variables include aspects within certain parameters of context validity that are either extra or unrepresentative to the constructs being tested e.g. more time allocated to test tasks or '…the use of language that is not entirely accessible' (Young, 2008, p.4).

Other challenges and difficulties with context validity include (1) "…the difficulty we have in characterizing language proficiency with sufficient precision to ensure the validity of the representative sample we include in our tests, and (2) the threats to validity arising out of any attempts to operationalize real-life behaviours in a test" (Weir, 2005, p.20).

Simply expressed: "The issue here is the extent to which the test content forms a satisfactory basis for the inferences to [be] made from test performance" (McNamara, 2000, p.50). This view is supported by Weir et al. (2009): "If test task performance is to be used to support inferences about performance in the wider domain of real-world tasks it is essential that both target reading activities and test tasks be described in terms both of cognitive processes and of contextual parameters" (Weir et al., 2009, p.106). It follows from this statement that, studying the contextual parameters of the test tasks in the Foundation Programme in Oman should help to establish their validity as representative of tasks involved in reading at academic level. This should also reveal whether there is construct under-representation or construct irrelevant variance involved in these tests with implications for validity. It is important to determine that the sampling of tasks based on the syllabus is sufficiently representative of the skills and abilities and the outcomes that were expected in the syllabus, as these, in turn, are expected to be representative of real life performance.

It is for all these reasons, in addition to the overall clarity of Weir's (2005)

framework, further developed by Khalifa and Weir (2009), that this framework is a very practical template for identifying the variables that need to be tested. These are thoroughly discussed in the next section.

2.7 Contextual features and the variables that affect the nature of reading

2.7.1 Overview

The current research context focuses on the validity of inferences being drawn from test scores regarding the ability of the test taker to be able to adequately comprehend academic texts. In devising test tasks, it is assumed that these tasks "… reflect both the cognitive processes engaged by the reader and textual features of reading tasks encountered in the wider domain of real-world tasks to which test performance is intended to generalize" (Green et al., 2010, p.192). In the previous section, it was noted that Bachman and Palmer (1996) had argued for test tasks to be as authentic as possible in situational and interactional contexts which tested the cognitive processes required for academic reading and the textual features that test takers would encounter in academic reading. This closely resembles the contextual and cognitive validities that were examined in Weir's (2005) framework. This framework was further developed by Khalifa and Weir (2009).

Alderson (2000) classifies the features involved in testing reading into two categories: the first comprises what research has found in relation to factors within the reader; the second comprises the features which research has found in the text which is to be read. This also echoes the earlier model of Bachman (1990), and the work of (e.g. Cohen, 1994).

First, the features that affect the reader's ability to read are examined. The chief of these is the state of the reader's knowledge and a second variable is the reader's motivation to read and how this interacts with why the reader is reading the text at all (Alderson, 2000). The importance of motivational factors in the Omani context is recognised and is briefly commented on at a

later point, although a full treatment lies beyond the scope of this study. However, writers such as Weir (2005) and Bachman and Palmer (2010) note the importance of motivational and affective factors in second language assessment. Csapó and Nikolov (2009), in their longitudinal study of second language proficiency, have highlighted the importance of affective factors in the development of proficiency in a foreign language.

Another characteristic concerns the strategies that readers use when processing texts (Anderson, 1999a; Alderson, 2000). Also important are relatively stable characteristics of readers such as sex, age and personality, which were pointed out by (Samuel and Kamil, 1988 and Brown and McNamara, 1992). Green and Jay (2005) also argue for attention to be given, at the pre-editing stage of test design, to features such as topic, topicality, level of language, suitability for the task, length, focus of text, style of writing, focus of task, level of task (Green and Jay, 2005).

These have been studied side by side with physical characteristics such as eye movements, speed of word recognition, automaticity of processing (e.g. Grabe, 1991).

A more recent development for context validity is found in Khalifa and Weir (2009), and which was also an important basis for the following studies on reading in second language: Weir et al. (2009), Green et al. (2010), Green et al. (2013). Khalifa and Weir's framework lists the following features for context validity in Table (2.3):

Table 2.3 Context validity features

Context validity	
Task setting • Response method • Weighting • Knowledge of criteria • Order of items • Channel of presentation • Text length • Time constraints	**Linguistic demands** **Task input and output** • Overall text purpose • Writer-reader relationship • Discourse mode • Functional resources • Grammatical resources • Lexical resources • Nature of information • Content knowledge

Adopted from Khalifa and Weir (2009, p.82)

The features are divided into task setting and linguistic demands. Each of these is examined in the following subsections.

2.7.2 Task setting

Early conventions in the history of language testing show the importance of different aspects of test task setting:

“The test sheets are printed on paper of a uniform size and bound with appropriate record sheets in pamphlet form. Exactly the same tests are given to all grades from the fourth through high school and university, and the conditions may be kept uniform, instructions for giving, scoring, and tabulating the test have been printed in convenient form together with necessary record sheets, answer cards, and graph sheets” (Courtis, 1914, p.376). Since then test task setting has considered a number of other important features which have been listed in Khalifa and Weir (2009) (Table 2.3 above) and these are discussed in the following subsections.

2.7.2.1 Response method

The literature suggests different test response methods measuring different aspects of language ability (e.g. Kinsch and Yarborough, 1982; Spiegel and Fitzgerald, 1990; Weir, 1990; Kobayashi,1993; Alderson et al., 1995; McNamara, 2000; Hughes, 2003; Weir, 2005; Fulcher 2010; Green, 2014).

Selected responses (e.g. multiple choice) or constructed responses (e.g. write a short paragraph), is the clearest organising principle for response method in reading (Khalifa and Weir, 2009). Presented below are the response methods that are most commonly utilized in the Omani second language test in the FP.

A. Selected response format

Multiple Choice Questions (MCQ): Well-constructed MCQs are useful for discriminating between strong and weak students (Hughes, 2003; Khalifa and Weir, 2009; Weir, 2013). Additionally, the level of difficulty can be easily increased or decreased through careful selection of texts and manipulation of distractors. They are appropriate for large-scale assessments and testing detailed understanding. However, Khalifa and Weir (2009), and Weir (2013) express some concern about MCQs for assessing higher-level processing. In fact, Rupp et al. (2006) go even further and question their value at all for higher order comprehension. The subjective nature of designing questions testing comprehension has been pointed out by Bernhardt (1991) who asserted that “inter-rater reliability is rarely high since there is general disagreement on the determination of the “most important parts of the text” (Bernhardt, 1991, p.199). This type of response can become merely a process of problem-solving which depends on verbal reasoning rather than integrating propositions (Rupp et al., 2006). A note of caution is issued by Valette (1977) who, in considering the problem of passage independence and prior knowledge, counsels against tests becoming exercises in logic and problem-solving rather than testing reading. Another shortcoming of MCQs is the fact that they may not validly test a real life reading situation (Freedle and Kostin, 1994; Khalifa and Weir, 2009). It is more effective when the test taker

first reads the text and decides the answer from the text and only then looks at the options (Khalifa and Weir, 2009). MCQs also can employ discernment skills as the reader must decide which type of reading, either expeditious or careful, is appropriate for each question. Another concern about MCQs is also posited by Weir (2013), "…that the mental model which is normally created while reading a text is affected if candidates try to incorporate all the options provided in an item into an ongoing text representation" (Weir, 2013, p.157).

Another type of MCQ is ***True/False items***, which is suitable for lower level reading. Khalifa and Weir (2009) also make the point that this type of MCQ allows the widest sampling content per unit of testing time. They also point out that the reliability of a test tends to increase the more that items of this type are included in the test. It is appropriate for assessing the use of scanning skills (expeditious reading at local level) but not necessarily for higher level i.e. skimming or searching (implied in the Matrix Table 2.1 above).

B. Matching

***Multiple matching*:** This type of test is useful for assessing the ability of the test taker to make appropriate use of the redundancy of language in order to locate the answer. Some of the options may come very close to answering the question but this is a matter of selecting that piece which captures every important element from the original text without omitting any material element. "In making use of redundancy, the reader makes use of prior knowledge, using something that is already known to eliminate some alternatives and thus reduce the amount of visual information that is required. Redundancy represents information you don't need because you have it already" (Smith, 2004, p.65). Cohen (2012) highlights another type of redundancy and elimination, a strategy that is performed in language tests namely, 'test-wiseness strategies': "…with regard to a reading test, it would mean using the process of elimination rather blindly (i.e. selecting an option without really understanding it at all, but rather out of a vague sense that the

other options are not likely to be correct), using clues in other items to answer an item under consideration" (Cohen, 2012, p.97).

C. Constructed response formats

C.1 Short answer questions (SAQ): These test expeditious reading at both global and local levels i.e. skimming for gist and searching for main ideas. Khalifa and Weir (2009) claim that there is greater certainty in these types of questions that a correct answer is due to comprehension rather than any other reason. A possible disadvantage of SAQs is that "…they involve the candidate in some writing and there is some concern that this interferes with the measurement of the intended construct" (Khalifa and Weir, 2009, p.67). Thus, there are implications for validity. Indeed Khalifa and Weir (2009) urge test setters to take care that there can only be one possible acceptable response when setting a test task. Also, this would call for special attention on the part of the marker in the interest of fairness. So, it runs the risk of marker unreliability (Khalifa and Weir, 2009). This argues for the need for marker training prior to marking the test.

C.2 Random deletion close and selective deletion gap filling:

Random deletion simply means, for example, deleting indiscriminately every fifth word whereas selective deletion involves an intentional deletion based on, for example, verbs. There is much controversy surrounding these types of tests (Brown et al., 2012). The average first language English speaker uses about two thousands word in day-to-day conversation. The educated L1 speaker still uses about 2,000 words in day-to-day conversation but has knowledge of a further 20,000 (Goulden, et al., 1990). Moreover, it was suggested by Huckin et al. (1993) that, knowing the meanings of the most frequent 2,000 words in English would be adequate for knowing over 80% of words on many pages of text. According to Khalifa and Weir, a 20 item test of any vocabulary would only sample 1 word in 500 (0.2%) from a 10,000 word vocabulary. This raises the question of how scores on this task can be

generalized to how the candidate might cope with broader demands on lexical knowledge (Khalifa and Weir, 2009), who also cite the works of Alderson (1978); Kintsch and Yarborough (1982) and Markham and Kobayashi (1995), as further evidence.

The principal criticism lies in the fact that close procedures are more reliable at local than at global level: “Close tests or selective deletion gap filling do not necessarily reflect the reader’s ability to comprehend beyond the sentence” (Khalifa and Weir, 2009, p.89). The weight of all this evidence amounts to the disconfirmation of earlier theories of a unitary understanding of language assessment such as was proposed by Oller (1979). Purpura (1999) reports “Oller’s (1979) claims generated several empirical studies on the nature of second language proficiency and eventually his notion of second language proficiency as a single unitary trait was disconfirmed in favour of a multi-componential view of second language proficiency proposed by Bachman & Palmer (1982, 1983)” (Purpura, 1999, p.17). A similar critique was also reported by Kunnan (1995). Close or single gap filling tasks measure lexical access and syntactic parsing skills. Furthermore, scores on these tests may only provide limited information on reading ability and would not reliably assess whether the test taker was a competent reader. Bernhardt (1991) draws attention to the fact that cloze testing only measures reading within clause boundaries by focusing the reader’s attention on individual words to the neglect of a global comprehension. He further argues that identical scores were found to be generated by cloze passages presented randomly and concluded that this type of test task did not adequately assess the reader’s understanding of the coherency of a discursive text.

To conclude this subsection, the following simple table (Table 2.4) is presented as a guide to choosing the right type of assessment to link to classroom activities:

Table 2.4 Relationships between assessment types and corresponding classroom activities

Assessment types	Appropriate for assessing	Example items	Corresponding classroom activities
Selected-response	Knowledge of (vocabulary, grammar, sound contrast, etc.) or receptive skills of listening and reading	True-false, Multiple-choice, Matching	Explicit learning of receptive grammar, vocabulary, and pronunciation knowledge exercises; all kinds of listening or reading activities in isolation, etc.
Productive-response	Productive skills of speaking and writing or their interactions with other skills, or task performance	Fill-in, short-answer, task-performance, etc	Productive knowledge of grammar, vocabulary, and pronunciation exercises; pair work, group work; role plays; speaking and writing performance tasks of all kinds, etc.
Personal-response	All four skills simultaneously, or higher-order thinking skills, or for motivating students to speak or write	Self-/ peer-assessments, portfolios, conferences, etc.	All of the activities immediately above, plus introspection and reflection activities, individualized instruction, project work, etc.

Adopted from Brown (2012, p.135)

Finally, having a wide variety of types of tasks has been highly recommended as it provides multiple ways for learners to provide evidence of their strengths (ETS, 2009).

2.7.2.2 Weighting

Various tasks in a test can be assigned different maximum scores (Hughes, 2003; Weir, 2005: ETS, 2009). This is based on a belief of test designers that certain items are more important than others and should therefore carry more weight in scoring. However, Alderson et al. (1995) point out that differential weighting does not often result in increased reliability or validity. Ebel (1979) cited in Alderson et al. (1995, p.149) had previously argued that reliability and validity could be better improved by assigning to a certain area, deemed to be twice as important as another, twice as many items rather than equal numbers of items compensated for by increasing the scoring.

Khalifa and Weir (2009) highlight the importance of the candidate being informed in advance of any weighting being employed so that they can monitor their allocation of time accordingly. O'Sullivan (2012) also highlights the importance of candidates knowing in advance the weighting of different tasks as impacting on goal setting.

The second point Khalifa and Weir (2009) make concerns the need for a sound rationale for the allocation of different weightings (e.g. more marks being awarded for skimming than for scanning). The argument amounts to whether to apply weighting in the scoring or to score all items equally but for greater representation for more important items.

It would seem however, in practice, that some element of weighting is unavoidable in order to make allowance, e.g. for situations where a certain test task might take more time as it might take longer to read than another test task.

The decision of the candidate as to how much time or energy to allocate to items with different weightings is a function of discernment (Weir, 2005) related to the optimal use of time. The fact that this involves a high level

ability of strategic decision-making implies that invigilators should give clear guidance prior to the test and that weightings are included on the test papers. Furthermore, in teaching and examination preparation, there should be sufficient practice of scenarios where decisions have to be made based on weightings. Instructions given immediately prior to or even during a test related to weightings are insufficient. Test takers need preparation well in advance so that they can make prudent decisions on allocation of time and time management generally. In the context of the Omani colleges, more attention needs to be given in the early preparation for testing by a thorough explanation of how test tasks are relatively weighted and practice should be given to strategies of time allocation based on these weightings. But perhaps even more effective would be for test designers in Oman to have a sound rationale for weighting and if possible to opt for greater representation of more important tasks rather than employing a weighting system.

2.7.2.3 Knowledge of criteria

If anything other than reading and comprehension is taken into account in scoring, this must be stated in the criteria by which the work is assessed (e.g. if spelling, punctuation…etc. are being taken into account the candidate needs to know). In such a case, writing skills are muddled up with reading and comprehension (Khalifa and Weir, 2009). Weir argues that the inclusion of certain criteria used in marking affects both "*planning* and *execution* mechanisms in the cognitive processing involved in task completion" (Weir, 2005, p.63). Both teachers and students need to know the criteria being used to assess and score and these should be known well in advance of the examination (Weir, 2005). Mustafa (1995) found that Arab undergraduate students considered that having prior knowledge of the testing criteria coupled with preparation practice were important for enabling them to show their true ability in the various test tasks. His study showed that students who participated in a preparation course for written assignments had a significantly higher percentage (42%) of achievement rate compared with those who had not participated in the course (23%) (Mustafa, 1995). This has clear implications also for preparation in reading and comprehension

tasks. Students clearly have a greater opportunity to show their true abilities by a correct understanding of the assessment conventions.

A fortiori, the American Psychological Association National Council on Measurement in Education (1999) cited in Khalifa and Weir (2009, p.63), highlights the rating criteria as being one of the most important pieces of information that a candidate should have prior to taking a test. In fact, the ETS Standard for Quality and Fairness (ETS, 2014) declares that candidates have a right to such information and Cambridge ESOL see the provision of such information as impacting on validity (Khalifa and Weir, 2009) - what Kunnan (2004) describes as 'condition familiarity' under the quality of access for fairness.

In the light of these findings, assignment and test preparation courses should be systematically planned and incorporated into the curriculum in Oman.

2.7.2.4 Order of items

For careful global type of reading, the representation of what is read happens incrementally, i.e. each sentence adds something to what has been established in earlier sentences so that the meaning gradually unfolds. Clearly, the order of items in a test in this situation is important (Khalifa and Weir, 2005). With expeditious types of reading, order of items is less important. In fact, where scanning is involved it is preferable and fairer to randomize the order as asking the questions in sequential order defeats the purpose of scanning which is to search the whole passage quickly for key words or concepts (Weir, 2005). In summary, where the cognitive process depends on logical sequencing and where each sentence reveals a new aspect what is being discussed, it is clear, in such a situation, that the order of items is important and should be logically sequenced. Alternatively, for expeditious reading a random order is probably better as the cognitive skill being tested is scanning.

2.7.2.5 Channel of presentation

Research suggests that comprehension is aided by information presented in more than one form (Larkin and Simon, 1987; Robinson et al. 1998; Kauffman and Kiewra, 2009). For example, text accompanied by a diagram, a picture or chart aids working memory (Hegarty, 1991, p.66 cited in Khalifa and Weir, 2009), as working memory is limited in holding a significant amount of complex information. Real life situations involve the use of more than one channel, for example, instructions on how to wire up an electric plug which would be much more difficult if either text or diagram alone were used. In the content of the current research, the use of multiple-channels in testing is authentic as in Colleges of Technology many textbooks use both texts and a chart or diagram to convey information. A good example of a test task based on multiple channels is to be found in the Canadian Academic English Language (CAEL) test which includes a reading test task requiring the test taker to make a selection from a number of words and use these in labeling a diagram or chart or to complete a table (Malone, 2010). This is a good example of using cognitive skills in comprehension of information that is provided by more than one channel.

2.7.2.6 Text length

If it is intended to measure the ability of the test taker to judge the relevance or irrelevance of certain details or to distinguish between main points or subsidiary main details, according to Alderson (1996) cited in Khalifa and Weir (2009, p.99), a long text is needed for these operations to be truly and realistically tested. Determining how long a text needs to be is not a straightforward task and most authors avoid being too prescriptive about the text length. The actual length depends on the purpose of testing, e.g. Test of English for Educational Purposes (TEEP) used texts of over 1,000 words on the grounds that texts of such length were more representative of real life situations than the texts found in IELTS and TOEFL (Weir, 2005). Khalifa and Weir (2009) provide a set of guidelines for test designers indicating the number of items and the total number of words involved in those different items in (Table 2.5):

Table 2.5 Text length in Main Suite Reading Papers

Examination	Overall number of words	Number of texts	Maximum for any single text
KET (A2)	Approximately 740-800 words	4	250
PET (B1)	Approximately 1,450-1,600 words	5	550
FCE (B2)	Approximately 2,000	3	700
CAE (C1)	Approximately 3,000	6	1,100
CPE (C2)	Approximately 3,000	9	1,100

Adopted from Khalifa and Weir (2009, p.101)

Spyridakis and Standal (1987) found a significant effect between passage length and comprehension. Nuttall (1996) cited in Khalifa and Weir (2009, p.99), sees the need for a longer text to test skimming and scanning. “In general, the longer the text candidates are presented with, the greater the language knowledge that might be required to process it. If short texts are not making the demands on these resources that will occur in normal cognitive processing, theory-based validity is compromised” (Weir, 2005, p.74). Cushing-Weigle (2000) comments on the impact on instruction and curricula of The Michigan English Language Assessment Battery (MELAB) reading section, which relies on short reading passages with multiple-choice questions. According to Cushing-Weigle, this does “…encourage reading strategies that are not always applicable to academic reading in practice. Most of the items depend on comprehension of a single sentence within the passage rather than the passage as a whole” (Cushing-Weigle, 2000, p.453). Bernhardt (1991, p.193) cited in Cushing-Weigle (2000), makes the point that “many L2 readers are able to deal with ‘units of language as separate entities’ but this does not necessarily lead to comprehension of a coherent message within a text” (Cushing-Weigle, 2000, pp.452-453). In conclusion, the view that a long passage of text should be included is well-supported for

testing cognitive skills based on coherence and cohesion within a text. Short texts may not test such cognitive skills where the comprehension of the passage as a whole is involved rather than finding individual pieces of information in a shorter text.

2.7.2.7 Time constraints

Mosback and Mosback (1976) encouraged speed reading in their book entitled '*Practical Faster Reading'*, and devised the following scheme for measuring word speed (Table 2.6):

Table 2.6 Measuring word speed

Reading time (min/secs)	Speed (w.p.m.)	Reading time (min/secs)	Speed (w.p.m.)
1.00	500	3.10	158
1.10	427	3.20	150
1.20	375	3.30	143
1.30	334	3.40	137
1.40	300	3.50	131
1.50	273	4.00	125
2.00	250	4.10	120
2.10	231	4.20	116
2.20	215	4.30	111
2.30	200	4.40	107
2.40	188	4.50	104
2.50	174	5.00	100
3.00	167	/	/

Adopted from Mosback and Mosback (1976, p.ix)

Alderson cautions against comprehension that is not time bound: “Speed should not be measured without reference to comprehension, but at present comprehension is all too often measured without reference to speed” (Alderson, 2000, p.30). Authenticity demands that the cognitive skills involved in comprehension also need to be time constrained as, in the real world of academic reading, time will likewise be limited. Various authors have measured reading speed in terms of words per minute (w.p.m.). In fact, Anderson (1999a; 1999b) states that there is little agreement on what the optimal reading rate is. Anderson (1999b, pp.59-60) summarizes some of the different views of the subject:

- Higgins and Wallace (1989): 180 words per minute may be a threshold between mature and immature reading, below this is too slow for efficient comprehension or for the enjoyment of text
- Dubin and Bycin (1991): 200 words per minute would be absolute minimum in order to read for full comprehension
- Jensen (1986) suggests 300 words per minute is the optimal rate
- Nuttall (1982, p.36): for a L1 speaker of English of about average education and intelligence …the reading rate is about 300 w.p.m. The range among L1 speakers is very great; rates of up to 800 w.p.m. and down to 140 w.p.m. are not uncommon.

Anderson (1999b) emphasizes the value of rapid reading for L2 students provided that it does not result in a decrease in comprehension. In fact, the results of an experiment by Chang and College (2010) set in Taiwan, showed that an experimental group with a timed reading intervention showed an improvement of 25% in reading speed and comprehension. Thus, they recommended the inclusion of a timed reading activity on a weekly basis in the curriculum stating that even 15 minutes practice could improve a learner’s reading speed and confidence.

In recent literature, with slight variations, the consensus seems to focus on around 150 (w.p.m.) as the measure for a slow reader: 250 (w.p.m.) for a fair reader and 350 (w.p.m.) for a good reader. Readers regarded as very superior would have speeds in the range of 300-600 (w.p.m.) (Khalifa and Weir, 2009). Chinese L2 readers of English are reported as having a mean

rate of 86.5 (w.p.m.) with a comprehension average of 63.9% (Haines and Carr (1990) cited in Khalifa and Weir, 2009, p.102). These compared with American L1 readers whose mean reading speed was 254 (w.p.m.) and comprehension mean was 75.3%. In view of Alderson's statement above, comprehension without the constraint of time does not validly measure expeditious reading (Weir, 2005; Khalifa and Weir, 2009), a skill that is of crucial importance for academic reading and comprehension. Equally, reading speed alone does not tell much about reading ability unless there is also a measure of comprehension of what has been read. Since, for the purposes of this type of test, the reader is informed that there is a strict time constraint, the activity of discernment is therefore required just as it would be in the real life academic situation.

A good example of a test based on speed and comprehension is given in the College English Test (CET) reading section in China, which involves a number of time-constrained tasks such as in-depth reading of 3 passages with multiple choice answers in 25 minutes and skimming and scanning one or two long articles of approximately 1000 words in 15 minutes (Zheng and Cheng, 2008). Here we see how comprehension and speed are being rated together. However, in view of what has already been discussed regarding the discernment function, it really ought to be a strategic choice on the part of the test taker as to which type or permutation of careful or expeditious reading at global or local levels should be employed in a given situation. This raises the problem of test task setting where a passage is chosen to test, for example, expeditious reading but it could be the case that in certain situations individual students might have opted for a more careful reading of a text whose content appeared to be very unfamiliar to them. It clearly cannot be the case that comprehension can be, in all situations, a function of time and speed as there are many situations where careful reading would be more appropriate for a given student in a given situation. A practical solution is to incorporate a number of items into the reading of a given passage, some of which test expeditious reading and others which call on the reader to read a certain number of sentences very carefully in order to answer a more subtle question.

However, in practice in the CTs in Oman, there is no provision for a time-constrained test as there is only a single passage (approximately 500 words) from which approximately 26 test items related to different types of tasks are to be completed within an overall time frame (see Appendix 2). In other words, it is left to the individual test taker to decide how much time should be allotted to each item. Clearly the issue here is not about individual tasks being time-bound but rather allowing the test taker to exercise some discernment in their allocation of time to individual tasks.

In the Omani context, there is a need for adequate practice of both types of tasks, constrained and unconstrained by time. This is supported by Nation, who advocates ample practice sessions in increasing silent expeditious reading speed (Nation, 2005) as a strategy that was also found to be successful in the Taiwanese study cited above. Therefore, heads of language centres in the Omani colleges are urged to build in practice sessions in classes well in advance of the actual exams so that students are given every opportunity to practise both types of tasks. Additionally, heads of centre need to ensure that the mock exams reflect both types of situations and that students are given adequate feedback to address any misunderstandings or inappropriate divisions of their time.

2.7.3 Linguistic demands: Task input and output

The second category for context validity is that of linguistic demands (see Table 2.3 Section 2.7.1 above). This is based on the communicative approach which began in Hymes (1972) and was followed up by Canale and Swaine (1980), Canale (1983), Bachman (1990), and Bachman and Palmer (1996). Linguistic demands need to be as close as possible to real life language use which, in the case of this research, is a college of technology.

2.7.3.1 Overall text purpose

In assessing reading, Alderson (2000) asserts that "…the text on which the assessment is based has a potentially major impact on the estimate of a

reader's performance and ability" (Alderson, 2000, p.255). A further important reflection to be borne in mind is the contention that "…since purpose and task both relate to the choice of text, a consideration of text type and topic is crucial to content validity" (Alderson, 2000, p.255).

Cushing-Weigle (2002, pp.8-10), based on the general model of writing discourse by (Vähäpässi, 1982), presents a model in which text types are divided into two categories; cognitive processing and main purpose. The main elements of overall purpose are based on:

- Referential (intended to inform)
- Conative (intended to persuade)
- Emotive (intended to convey feelings or emotion)
- Poetic (intended to entertain, delight, please)
- Phatic (intended to keep in touch)

Texts, in the past, have been mainly referential in type but Khalifa and Weir (2009) have drawn attention to the recent increase in phatic and emotive written communication in blogs, text messages and chat rooms. Nevertheless, the only implication for validity here is that test tasks relate to overall purpose. For example, in referential texts the tasks must be confined to points of information that are explicit or perhaps implicit in the text.

2.7.3.2 Writer-reader relationship

The intended reader or audience is of crucial importance in the creation of a text and its meaning. Hill and Parry (1992) point out that "…writers do not simply refer to the world they are writing about; they associate themselves with particular communities of language users" (Hill and Parry, 1992, p.445). Ede and Lunsford (1984) cited in Khalifa and Weir (2009, p.106) distinguish between two types of audience: the audience addressed and the audience invoked. The audience addressed is the intended reader whereas the audience invoked may be a fictitious reader which often happens where the writer is writing for a rhetorical purpose (Khalifa and Weir, 2009). The intended audience determines the extent of content knowledge that the writer

can assume that the reader already has. It is postulated that, in a socio-cognitive view of L2 reading, some readers tend to see different things in the same text. This results in individual decisions regarding what is important for them in the text and explains how they make sense of it or reconstruct it. Thus, Bernhardt (1991) is able to assert that "…the input text and the output text are, in this integrated view, different entities" (Bernhardt, 1991, p.15). Thus, the processes of the texts presuppose the knowledge that the reader already possesses. However, this raises issues relating to validity as it hinges on the assumptions made about what the reader already knows, for example, the target age, range and background features (what Bachman (1990) refers to as test taker's characteristics). Khalifa and Weir (2009) comment on the amount of scrutiny that goes into the pre-editing of test tasks. Nevertheless, it must be a daunting task to ascertain what knowledge of the world different groups can be assumed to possess. For example, IELTS reading texts often present a section of a scientific report on, for example, the environmental impacts of certain phenomena or human behavior. The assumption here is that the test taker has sufficient knowledge to already have an awareness of impacts on environment to be able to understand that new information, supported by a graph or table, should enable them to comprehend the overall meaning of the text in the light of the knowledge possessed by the test taker. Of course, a test taker could, by chance, have a profound knowledge of the subject which would therefore be an advantage to them over other test takers. However, the variety of topics used for comprehension must be assumed to balance this out as the next item might relate to sport, of which that test taker might have very limited general knowledge. Nystrand (1989) cited in Weir et al. (2009, p.106) makes the point that meaning is created in an inter-subjective way between the participants of a discourse and resides in the expectations and assumptions that both have of each other. The cognitive ability being tested here is schema-related in that meaning in a written text is created through drawing on an assumed prior general knowledge of the subject at issue.

Writing, rather than being an isolated individual action, involves the endeavours of both the reader and the writer and is shaped through mutual

assumptions involved in the understanding of rhetorical situations (Weir et al., 2009).

Consequently, designing tests for the Omani situation implies that the test designers are cognisant of particular assumptions that can be made regarding what is already at least partially known by the reader and those aspects of the tests which may be outside the range of their knowledge. Simply stated, this means that test designers must know the intended target audience for whom the test is being designed. In the Omani context, it can be assumed that, due to the age of the test takers, they have knowledge and experience of 'social media' and would probably have Facebook, Twitter and Instagram accounts. Furthermore, through the internet social media and films, they would have some knowledge of global culture; however, they might have little or no experience outside of Oman and test designers would need to be careful in making assumptions, for example, that they have any understanding of the processes involved in travelling abroad such as obtaining visas and planning itineraries.

2.7.3.3 Discourse mode

It is acknowledged in the literature that an understanding of how a text has been organised can greatly aid the reader's comprehension (e.g. Norris et al., 1998; Lacroix, 1999; Anderson, 1999b; Degand and Sanders, 2002; Farrall, 2012). In discourse mode, an argument or point of view or discussion is gradually built up in a logically or chronologically coherent way. Once the reader discerns that he/she is reading a discursive piece of text, they should be able to quickly locate the key idea on which the entire paragraph hinges. A reader might expect the key point to come in the conclusion which is likely to be the final or penultimate sentence. However, this is not always the case. The key point might occur early in the paragraph but might also require additional sentences to add coherence to the claim being made.

Kaplan and Grabe (2002) identified four distinct strands in the research into the contribution made by understanding the nature of a written text: (1) *Textlinguistics*, which sees a text as 'a stretch of language' where the

textuality is derived primarily from the cohesion and coherence inherent in the text, for example, how the sentences are linked together, (2) *Cognitive*, referring to the psychological processes involved in comprehension. Discovery of meaning is a communicative event and where meaning is not seen as entirely residing in the text, (3) *Discourse analysis*, which refers to the writer-reader relationship and embraces the specific context and the writer's intentions and also involves the use of genres, and (4) *Contrastive discourse analysis*, which also embraces issues of meaning that arise in translation to another language. However, the authors do not suggest that this is a strict classification scheme but rather that an understanding of the comprehension of written text more likely draws on all of these strands.

Of the different discourse modes, studies show that comprehension is facilitated to a greater extent by problem-solution mode, comparison mode or causation structure than other types of modes such as description (Khalifa and Weir, 2009).

Discourse mode includes a number of devices, many of which call for higher level or even more nuanced cognitive skills. They include cohesive devices. Here, the correct use of connectives between sentences is so important. The reader has to grasp the implications of 'moreover, however, consequently, in contrast...etc.' Goldman and Rakestraw (2000) and Enright et al. (2000) underline how cohesive devices establish textual coherence. However, Weir et al. (2009) assert that the effect of cohesive devices on comprehension may not be as clear-cut. Such a view re-echoes that of (Hudson, 2007): a L2 reader's "…formal knowledge of how cohesive markers operate will affect the ease with which the text is processed and consequently will affect the reader's level of text comprehension. Additionally, knowledge of the system through which cohesion is established can assist second language readers when they confront trouble in text comprehension. The reader can explicitly examine troublesome text to reestablish the cohesive thread within the text" (Hudson, 2007, p.178).

Rhetorical features refer to how the text is constructed in a certain way, for example, problem-solution, comparison-contrast, or causation-effect

rhetorical structures often result in better recall and comprehension than other types of rhetorical features such as classification or description structures. This point is evident in many texts which involve, for example, classification where the writer is aware of the difficulty of comprehension. This also applies where the text is supplemented by a table or chart of classification or else the presentation of the items being classified is achieved through bullet point format. Likewise, description structures are often accompanied by drawings or photographs (Carrell 1984; Goh, 1990).

Freedle (1997) adds that finding the main idea in a text that is high in coherence is easier than is the case with a text lacking coherence. Weir et al. (2009) argue that the use of rhetorical features should be an important consideration in the selection of texts for tests of academic reading.

On the effect of rhetorical ordering on readability, Urquhart (1984) carried out experiments that provided evidence which showed that where overall meaning was understood as the goal, the reader was much better organized. For example, in the case of a test task which involved simplifying a passage based on its time sequencing, Urquhart advised that "…the reading teacher ought to be wary of exposing low-level learners to narratives which do not conform to time ordering" (Urquhart, 1984, p.174). The reason for this conclusion is that information presented in time sequence involves lower cognitive skills than for texts based on logical sequencing where a conclusion can be reached based on two or more separate assertions made in the text which leads to the implication that is asserted in the conclusion. This calls for high levels of cognitive skills in understanding how arguments and discussions involving contrasting opinions or evidence are being constructed.

Another classification of discourse modes is that of genre, which is seen by Swales as "a recognized event with a shared public purpose and with communicative intentions mutually understood by participants" (Swales, 1986, p.19). This refers to how comprehension is aided and assisted by the reader quickly recognising the particular genre of the text, for example, whether it is a historical account or a scientific report or a philosophical

argument. Norton and Stein (1998) as white researchers working in the context of redress in post-apartheid South Africa, experimented with a text about an 'encounter between monkeys and humans' with black students, to test a hypothesis that the test might be unfair on cultural grounds. Surprisingly, they found that 63% of test takers scored high (80% correct) (Norton and Stein, 1998). The assumption that genre types may lack validity on cultural grounds then is not always a correct one. Moreover, such an assumption has been thoroughly explored and critiqued in cross-cultural communication studies which emphasized the importance of the 'minimization of misunderstanding' in comprehension (Gudykunst, 2004).

However, Weir et al. take a somewhat more common-sense view of the importance of genre that it "…would seem logical to suggest that if texts to appear in a test are sourced from academic contexts they are likely to share lexical, syntactic and discourse features with texts encountered at a university" (Weir et al., 2009, p.108). However, there is some confusion or overlap of terms here regarding what is referred to as rhetorical task in Enright et al. (2000) which they divide into a three-fold classification:

- Exposition
- Argumentation/persuasion/evaluation
- Historical biographical/autobiographical narrative.

This seems to be remarkably similar to genre. For example, Cushing-Weigle's (2002) subcategories of exposition, on closer examination, again appear to fall under the umbrella of genre. In fairness, however, Enright et al. (2000) and Cushing-Weigle (2002) do broaden the understanding of genre by including finer distinctions between different discourse modes. For example,

- *Definition/description/elaboration*, providing full definitions of concepts, describe unfamiliar terminology, elaborate on terms specific to the discipline and clarify specific uses of the terminology
- *Illustration* involves providing examples or a short anecdote to fully describe an abstract concept.

- *Classification* involves grouping several items together according to similar features or principles, showing how discrete items belong to a larger group.
- *Comparison/Contrast* involves designating distinctions among concepts, particularly regarding their similarity and dissimilarity (Weir et al., 2009).

Urquhart and Weir (1998, p.14) cited in Khalifa and Weir (2009, p.109) urge test designers to provide evidence regarding which discourse modes are to be considered appropriate at the various proficiency levels. Khalifa and Weir (2009) point out two problems relating to this. Firstly, there is a lack of agreement on terminology among the writers and, secondly, there is the fact that so little research has been conducted into the level of difficulty of the task. Given that discourse mode and the impact of level of difficulty of a text are under-researched, it is clear that care needs to be exercised in this area regarding the validity of test tasks: “The crucial factor in selecting a reading text is whether the text would allow their intended reading activities to be measured” (Khalifa and Weir, 2009, p.111). They continue by citing some examples. Firstly, where the task is to extract the meaning of the text, some examples where there are sufficient main ideas are presented as well as examples of insufficient main ideas in the text. Secondly, for tasks testing inferencing skills, they cite examples of where sufficient information is given to be able to deduce the overall meaning and some examples where there is a paucity of information to such an extent that this would result in the task lacking validity (Khalifa and Weir, 2009). Clearly, to meet the requirements of validity, the test tasks must contain the requisite amount of main ideas or information. The implications of the foregoing discussion for testing in the Omani situation are to determine whether these different discourse modes, appropriate for level four, are sufficiently represented in test tasks.

2.7.3.4 Functional resources

“Function is a term used to describe the illocutionary force of what is said. Examples of communicative functions might be where a speaker has to persuade, advise, describe, etc.” (Weir, 2005, p.78). It is a device which is

not just dependent on the literal meaning of what is said but on its illocutionary force, what Bachman (1990) and Bachman and Palmer (2010) call 'illocutionary competence'. For example, at table, the question: 'Is the salad ready yet?' does not expect a yes/no informative answer but the prompt delivery to the table of a bowl of salad. The delivery of the salad represents the illocutionary force of what has been said. Earlier, Bolinger (1977) cited in Kaplan and Grabe (2002, p.199), had drawn attention to the fact that external facts pertaining to language use including cultural, writer intention and textual context, were inseparable from grammatical structure for functionality. A point is also asserted by Bachman and Palmer that "…functions do not normally occur only in individual, isolated utterances. On the contrary, the majority of language use involves the performance of multiple functions in connected discourse" (Bachman and Palmer, 2010, p.46). Understanding a discourse or comprehending a text entails the reader in drawing on many aspects including the *test taker's own characteristics* and those of other participants, *prior knowledge* and *knowledge of language use setting* (Bachman and Palmer, 2010).

Bachman and Palmer (2010) classify functional knowledge into language functions including, but not limited to:

- Ideational (based on real-life experiential knowledge),
- Manipulative (based on language used which affects the world around us),
- Heuristic (language for extension of knowledge through various devices such as problem-solving), and
- Imaginative (the use of language for fictional and figurative purposes).

Following the work of Wilkins (1973, 1976); Van Ek and Trim (1998, 2001) and North (2000) cited in Khalifa and Weir (2009, p.111), clear definitions of functional requirements began to emerge especially for the A2 to B2 levels in the CEFR (see also North, 2011). Hudson (2005) reported how various criteria of language functions were related to scalar measures on different theoretical and practical models including Bachman and Palmer (1996) and the CAEL. The earlier work of Bachman (1990), while most informative in

relation to levels of functioning, was, however, quite theoretical and needed to be translated into more practical terms.

Work is still needed to be done beyond level B2 in the CEFR and the English Profile Program carried out by Centre for Research in English Language Learning and Assessment (CRELLA). Green (2007) cited in Khalifa and Weir (2009, p.111) sets down some guidelines for developing the testing of functionality at C level in the CEFR. The CEFR implicitly suggests the qualitative changes in the functions that are expected of learners between A and B levels. “Level A and B appear to offer an expanding repertoire of communicative functionality reflected in expanding access to context for language use” (Khalifa and Weir, 2009, p.112).

Khalifa and Weir conclude that, at C level, there is a shift in functionality towards: “rational inquiry and exposition, argument and suasion” (Khalifa and Weir, 2009, p.112). In view of the need for further clarification, especially of terminology and the results of on-going research, it remains difficult to appreciate the implications of functionality for context validity. Nevertheless, it is still of importance to check in the Omani situation that different functional resources are adequately represented in test tasks in order to ascertain that the test taker is able to sufficiently differentiate between persuasive, descriptive, instructional, advisory and other types of communicative functions and how such differentiation affects the meaning that is to be ascribed to a piece of text or spoken word. However, herein lies a problem specifically for L2 readers. This lies in the very subtle functional language uses that belong to different languages and cultures. So, in the above example, the question as to whether the salad is ready or not is actually a polite way of presenting a demand for salad. In Arabic, in everyday use, it is more common to be direct so that, in a similar context, it is acceptable to simply say ‘please bring me the salad’. The implications here are that this is quite difficult to test especially for L2 readers and, as mentioned above, most of the writers comment on its underdeveloped nature as a test task. In fact, it appears to be a feature of careful reading at the global level with reference to Khalifa and Weir’s (2009) model. Accordingly, this would be, more typically, a feature for more advanced and able readers targeted for use in L1 situations.

It should, therefore, not be unduly represented in tests in the CTs in Omani situations bearing in mind that the students' career paths are targeted more towards the home market.

2.7.3.5 Grammatical resources

In the context of assessment, grammatical forms are important for inferring the exact meaning intended by the writer (Alderson, 2005). This section shows how over recent decades, a balance has been struck between an insistence on grammatical rectitude on the one hand and on communication of meaning on the other hand. Purpura saw the importance of grammar for "meaning in both content-impoverished (e.g. multiple-choice tasks) and context-rich (e.g. problem-solving tasks) test situations" (Purpura, 2004, p.63).

Alderson (2000) emphasised the importance of understanding syntactic structures in L2 reading and other authors such as Khalifa and Weir (2009) underline the important role that syntactic parsing plays in comprehension. The parsing of sentences was a common feature of the 1960s models of reading. Its value is once again being recognised but now in the context of meaning and understanding rather than analysis. Weir et al. (2009) based their study on empirical evidence provided by:

- Alderson and Clapham (1992) who pointed towards a very close relationship between a *test of grammar* and the *IELTS reading component*
- Shiotsu (2003) who found that syntactic knowledge played a central role in likely affecting performance in reading
- Shiotsu and Weir (2007) who evidenced the relative importance of syntactic over lexical knowledge in accounting for variance in tests of reading
- Berman (1984) who showed how opacity and heaviness of sentence structure were features encountered in academic texts: "…valid test of academic reading should reflect the syntactic features likely to be encountered in academic texts" (Weir et al., 2009, p.108).

The CEFR provides no guidelines for the grammatical range which candidates should be able to manage at various ability levels (Khalifa and Weir, 2009; Green, 2011). An attempt has been made by Alderson et al. (2004) to provide such guidance by suggesting four categories of grammatical complexity in relation to Cambridge ESOL. These are:

- Only simple sentences mapped to KET
- Mostly simple sentences mapped to PET
- Frequent compound sentences mapped to FCE
- Many complex sentences mapped to CAE and CPE

The Alderson et al. (2004) scheme is helpful in establishing the appropriateness of texts for proficiency levels and therefore serves as an aid for validity.

Another approach to grammatical complexity is based on sentence length, measured by the number of words. Weir et al. (2006) cited in Khalifa and Weir (2009, p.123), using the Flesch reading ease and Flesch Kincaid measures, found that IELTS reading tests were significantly easier ($p<.05$) to read than first undergraduate texts (Khalifa and Weir, 2009).

Table (2.7) shows the readability scores on a number of Cambridge reading tests when the passages were measured by the various readability measures:

Table 2.7 Difficulty estimates in Main Suite Reading papers

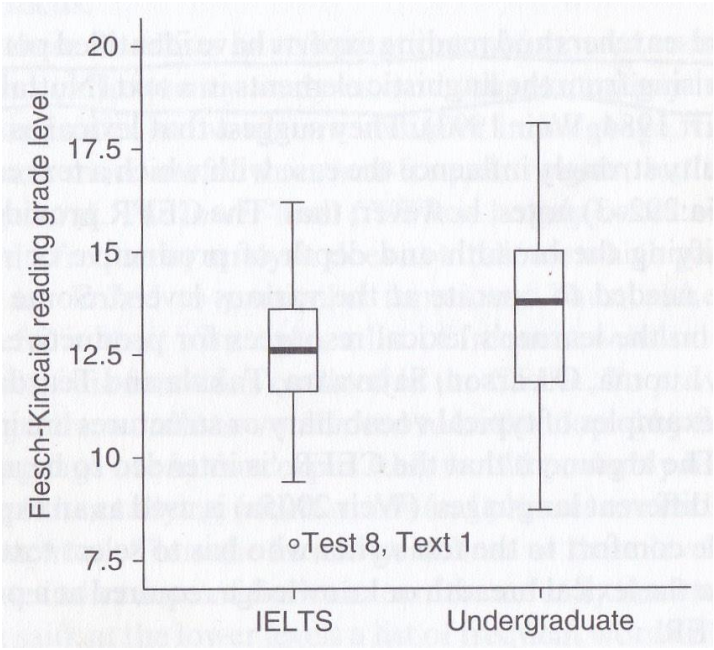
Main Suite Level	Flesch reading ease score	Flesch-Kincaid grade level	Flesh Kincaid range
KET (A2)	78.3	5.5	2-7.4
PET (B1)	64.7	7.9	5-10.1
FCE (B2)	66.5	8.4	5-12.3
CAE (C1)	58.4	9.6	5.7-16
CPE (C2)	57.7	9.9	5.6-16.1

Adopted from Khalifa and Weir (2009, p.122)

CAE (C1) and CBP (C2) were 9.6 and 9.9 respectively on the Flesch Kincaid scale whereas Weir et al. (2006) cited in Khalifa and Weir (2009, p.122) found that undergraduate levels on this scale should be around 13.5, thus, indicating that these two levels C1 and C2 were inadequate measures of the required ability to read at university.

However, these scales of text difficulty are not without their critics (Masi (2002) cited in Khalifa and Weir, 2009, p.122). Each of these estimates of text complexity was derived from either the average number of words in a sentence or the average number of syllables per word (Khalifa and Weir, 2009, p.122). Based on some combination of these two measures, applying the Flesch Kincaid reading grade did (as previously mentioned) show that undergraduates' average scores were a little higher than the average IELTS scores on this scale. However, this small difference proved to be significant. The matter is more complex however, when the range of scores is taken into account. The lowest score on IELTS is still higher than the lowest score for undergraduates see Figure (2.5):

Figure 2.5 Flesch-Kincaid reading grade levels of IELTS and first year undergraduate core texts



Aadopted from Khalifa and Weir (2009, p.123)

However, these results are based on the assumption that grammatical complexity can be measured in terms of words per sentence and/or syllables per words. In other words, syntactic knowledge can more than compensate for deficiency in lexical knowledge. Among other languages such as Persian, Chinese, Japanese, Portuguese and Danish, the English language is considered to be opaque in both directions, i.e. sound to symbol and symbol to sound, “making it a language that is both hard to read and spell” (Lems et al., 2010, pp.80-81). Opacity and heaviness increase the degree of difficulty in comprehending a text (Berman, 1984 cited in Khalifa and Weir, 2009, P.118). Such difficulty is explained in terms of parsing sentences— “…to recognize the basic constituents of subject-verb-object, noun-verb-noun relations, and so on” (Alderson, 2000, p.69). Nevertheless, the author hastens to add that it is “…unduly simplistic to believe that particular syntactic structures will always cause difficulty” (Alderson, 2000, p.69). Differences based on grammatical resources between Arabic and English

are noted in the literature: Arabic-speaking students, for example, generally speak one of their modern regional dialects, while a good portion of the written material is in Classical or Standard Arabic, which can be quite different from the spoken language used for all but the most formal occasions (Hayes-Harb, 2006; Taha, 2013). “While written English is also somewhat more formal than the spoken form, there is not such a wide divergence as in Arabic. Thus, students have to be taught to make the connection between the two” (Bruder and Henderson, 1985, p.9).

For Arab ESL readers, similar to the students in this study, opacity can be a hindrance. “Arabic is very consistent with almost a 1:1 phoneme-grapheme representation, and so the sound-symbol correspondences are relatively transparent. On the other hand, while English does have some consistent phoneme-grapheme representation, it also has inconsistent/more complex representations and so the sound-symbol correspondences are relatively more opaque than in Arabic” (Saigh and Schmitt, 2012, p.26). A suggested solution for English language learners is that they “must first develop the phonological awareness to perceive the (often subtle) differences in vowel sounds, and then, using probabilistic reasoning, try to match them with the grapheme or graphemes that seem most likely to go with the sound” (Lems et al., 2010, p.80).

The point is made that, in general, complex sentences are more difficult than shorter simple sentences. Nevertheless, if very elliptical language is used and colloquial lexis, short simple sentences will be found to be more difficult than longer sentences (Khalifa and Weir, 2009). A subordinate clause can often amplify further the meaning of the principal clause which, left to itself, might be somewhat ambiguous (Snow, 2002). This may help to explain why L2 readers sometimes find that comprehending articles in The Guardian is actually easier than the more colloquially phrased articles often found in the Sun or Daily Mirror. What these scales have done is to reduce grammatical complexity simply to word count or syllable count; in other words, a lexical measure of grammar.

An alternative approach to measuring grammatical complexity based on

other variables than simply on word per sentence or syllable per word is to be found in Cappelli (2010). Admittedly, he approaches this question from the perspective of a mathematical model based on dynamic complex systems in other fields but including language. The matter of the complexity of grammar was seen as consisting of many variables and was included in an overriding concept which he calls ‘texture’. However, he commented that texture was difficult to define as it was an emergent concept which depended on context (Cappelli, 2010); texture does, however, include more variables such as adjectives, verbs, nouns…etc., which would be more subtle indicators of grammatical complexity. A detailed outline of the actual method is, nevertheless, difficult to obtain, but is mentioned here to highlight other aspects of grammatical complexity.

An argument could also be made for including the number of subordinate clauses in a passage as a characteristic that might indicate grammatical complexity. This is supported by Purpura (2004) who asserted that grammatical knowledge embodied two closely related components. The first of these was knowledge of grammatical form which included linguistic categories such as lexical, cohesive and phonological. The second component was grammatical meaning which he otherwise referred to as “the literal meaning expressed by sound, words, phrases and sentences where the meaning of an utterance is derived from its component parts or the way in which these parts are ordered in syntactic structure” (Purpura, 2004 p.61). Thus, although Purpura is not concerned with the actual number of subordinate clauses as a measure of grammatical complexity, he has drawn attention to two important components in defining grammar for assessment purposes.

Current views of English grammar, represented by authors such as Huddleston and Pullum (2005) and Carter and McCarthy (2006) have underlined the importance of syntax and morphology (roughly speaking, sentence structure and word form respectively). Earlier Crystal (1966) had a similar division of grammar into morphology and syntax. In Carter and McCarthy's descriptive approaches to the learning and teaching of grammar,

they have proposed an incremental scheme. Such a scheme can be applied over the four levels in the Foundation Programme in Oman with an emphasis on syntax and morphology as the two principal elements of grammar.
At lower levels, it could begin with the simple sentence consisting of a single independent clause (Quirk et al., 1985) based on knowledge of subject-predicate e.g. simple sentences such as “Ahmed arrives.” or “Ships leave.”. Students can learn to move beyond such simple sentences to understanding how the use of noun phrases such as “Young Ahmed arrives.” or verb phrases such as “Ships leave early.” can show how simple sentences can be expanded. Then, students can learn how meaning is further enhanced by moving on to subject-predicate-object type sentences such as “Fatima eats bread” and then gradually building up to the use of adjectives and adverbs and linking signals such as “Well,”, “To begin with,” or “Besides,” within or between sentences (Leech and Svartvik, 2002). Students at higher levels in the Foundation Programme can move on to reading and comprehending multiple sentences that are both compound and complex (Quirk et al., 1985). For example, a compound sentence with two independent clauses such as “Go to lessons and pay attention.” can help students to use connectives to enhance meaning. Similarly, learning how to use a complex sentence such as “If you practice every day, you can become proficient in English.” can help the students to understand how meaning is enhanced through the use of one or more subordinate clauses. The incremental approach to grammatical complexity for pedagogic purposes is also helpful in guiding the measurement of grammatical complexity in L2 reading texts. Counting the number of subordinate clauses used in a passage might therefore be a useful indicator of grammatical complexity in input texts.

In his paper on assessing grammatical ability, Rimmer (2006) does not see any incompatibility between an approach based on grammatical meaning in a larger corpus and discrete measures of grammar based on lexis. In fact, he argues for including both for task setting. Although his focus was on grammar in general, his conclusions are nevertheless applicable to grammar in testing reading. Rimmer’s (2006) methodology actually combined the inherently subjective evaluation of grammatical complexity made by expert

intuition with input from corpora. This approach has been very effectively implemented in Green et al. (2010), where aspects of grammar in reading texts that were not directly amenable to scalar measurement were evaluated by experts, and this was combined with a more empirical approach to other grammatical features for which measuring instruments had been devised. Green et al.'s (2010) study adopted Khalifa and Weir's (2009) model on which this research is based.

In summary, Alderson observes that "the interaction among syntactic, lexical, discourse and topic variables is such that no one variable can be shown to be paramount. Moreover, even the ability to guess words from context has to be seen in context: the context of the reader, and other variables in the text" (Alderson, 2000, pp.70-71). A test of reading in Oman then, should include a wide range of the variables as indicators of the kinds of grammatical complexity that these students are likely to encounter during their post-foundation study and the predominance of one particular type should be avoided.

2.7.3.6 Lexical resources

Many authors, (e.g. Urquhart, 1994; Nuttall, 1996) comment on the influence of lexical difficulty as well as grammatical difficulty in influencing the ease of reading. Huhta et al. (2002, p.131) cited in Khalifa and Weir (2009, p.124) state that no examples are given in the descriptors for typical vocabulary or structures in the CEFR. Weir (2005) has explained this as arising from the CEFR's intention to be capable of application to a whole range of languages. Thus, there is little information in the CEFR as to the breadth of lexical knowledge that is required at various levels (Khalifa and Weir, 2009). Alderson et al. (2004) also point to the lack of definition generally in the CEFR; for example, the word 'simple' is used but without definition. They comment that the CEFR is generic in nature and that it cannot be expected to provide the detailed guidance that test writers require (see also Fulcher, 2010).

Vocabulary building has been approached in a number of ways in teaching

but especially through the use of word lists (Coxhead, 2000). Different schemes of word lists are in existence. Some are based on empirical studies (statistical analysis of frequency of use), more are somewhat intuitively based and some are functionally orientated (Laufer and Nation, 1995). For L2 reading instruction, Koda (2004) suggests equipping students with the 2,000 high-frequency words in the target language. It is claimed that "The logic is straightforward: The core vocabulary accounts for roughly 80% of the words in most texts (Nation and Newton, 1997), and therefore the sooner these words are learned, the better L2 reading comprehension is expedited" (Koda, 2004, p.59). However, Khalifa and Weir (2009) speak about the recent developments of computerized versions of written and spoken word lists and these can be broken down into words for L1 readers and words for L2 readers. Students who are progressing to university level need to know more than the two thousand words in common use in spoken communication. They need to build up their vocabulary to include not just technical words but also words used in discourse in building up an argument etc. These word lists are aimed at vocabulary building and are not just for memorisation but for use in various exercises and classroom activities such as gap-fill using these words so that the context of the word is also learned. Nevertheless, Csapó and Nikolov (2009) identify rote learning as one of the important components of language learning aptitude.

A more nuanced approach to the threshold theory of lexical knowledge is found in the recent work of (Ardasheva et al., 2012), who argued for a distinction to be made between vocabulary in an oral setting and in a more literate setting. Accordingly, a separate threshold is needed for either setting but in general the threshold level was not very different from that proposed earlier by Cummins's (1979, 2000) studies cited in Ardasheva et al. (2012, p.769). This seems to support a similar distinction made by Khalifa and Weir (2009) above.

Khalifa and Weir argue that word lists are useful for vocabulary building at lower levels and meet many of the requirements of functionality in the CEFR. However, at higher levels, they may be less useful (Khalifa and Weir, 2009) – of course this is true if word lists have already done their work at lower

levels.

In fact, two of the five predictors of lexical development in research conducted by Zareva (2005) were vocabulary size and knowledge of words from various frequency bands. These findings confirmed the earlier work of Henrikson's (1999) three-dimensional model (breadth, depth, and receptive-productive) and again showed that, as the learner progressed to higher levels of proficiency, these dimensions became less important as predictors. Admittedly, these were working within a paradigm of vocabulary acquisition based on trait theory which has been critiqued by Read and Chapelle (2001) who favoured an interactional approach for vocabulary testing. The point is that the trait theory assumes an underlying trait that is context independent whereas interactionalist approaches argue that learners' characteristics must be seen as relative to a specific context.

Khalifa and Weir (2009), within a socio-cognitive rather than a trait-based paradigm, suggested a threshold of 5,000 words that could be learned using exercises based on word lists (the threshold may differ from one language to another). The 5,000 word threshold for English represents words that are common to all academic reading; beyond 5,000 words the words become increasingly more technical and domain related. Another threshold suggested by Khalifa and Weir (2009) is the 3,000 words which make up the bulk of everyday conversation. However, they do argue that these number thresholds should be taken as "rules of thumb" rather than in a strict sense. In fact, they argue for higher thresholds than these, more like 5,000 word families, as the previous thresholds were based on 95% vocabulary recognition whereas, in the current view, this should be raised to 98% coverage. The argument is that having a vocabulary of around 5,000 words should enable any L2 reader to be able to cope with most academic writing. Unfamiliar words can be understood by inferring their meaning drawing on prior knowledge.

However, Widdowson cautions that "the criteria of frequency and range indicate the indexical value of a word, but they cannot be used exclusively for determining what is to be taught, quite simply because words of high

indexical valency are relatively empty of lexical content: they are auxiliary, enabling devices" (Widdowson, 1983, p.94) and takes the argument even further:— "In general, then, we may state it as a rule that the greater the lexical content of a word, the more narrow its indexical range: lexicality is in inverse proportion to indexicality. It follows from this that words of wide indexical range are especially useful for negotiating the conveyance of more specific concepts, for defining terms which relate to particular frames of reference" (Widdowson, 1983, p.93). It is clear that in the Omani context, for less able students, more emphasis might be placed on skills of inferring the meanings of unfamiliar words both from the context and prior knowledge as they are more likely to encounter unfamiliar words in the first year of their technology programmes. In fact, Hall and Durán (2009) have provided empirical evidence that lexical deficit can be compensated for by students who have developed the ability to infer meaning without the need to refer to their L1 and therefore learning to think in L2 is probably more important than simply focusing on vocabulary building. Thus, in the context of the FP students in Oman, strategies may need to be devised to help them to develop the ability to think in L2 and infer meaning without over reliance on L1.

Further implication for lexical resources is the need to minimise the use of highly technical words in the foundation programme test because these only begin to feature above the 5,000-word threshold mentioned above. Technical words related to their field of study actually require not a generalist definition for its everyday use but actually a strictly defined understanding of the term within that branch of knowledge. This is better left to the first year of their technology study or course.

2.7.3.7 Nature of information

This refers to the degree to which a particular text is posed in predominantly abstract or concrete terms. Khalifa and Weir (2009) argue that because concrete language very likely draws on cognitive operations of both verbal and non-verbal systems, it is therefore less difficult to comprehend than

abstract language which is restricted to the verbal systems only. For example, the degrees of abstraction involved in the word ‘chair’ could be ‘seat’, then ‘furniture’, then ‘furnishings’ and finally ‘entity’. With ‘chair’, ‘seat’ or even ‘furniture’ there are other cognitive processes at work other than verbal systems, for example, mental imagery of ‘a chair’. However, at the highest degree of abstraction, the word ‘entity’, is entirely dependent on verbal cognition or on what Moore and Morton (1999) call the ‘metaphenomenal’. In fact, drawing on the earlier work of Halliday (1994), Moore (2002) and Moore and Morton (2005) prefer to distinguish between phenomenal (real world situations, events…etc.) and metaphenomenal (theory, method, idea… etc.) to describe in more broad terms the nature of the information in respect of its concreteness or abstractness. Here, discernment is operant in deciding on the nature of the text in order to select the appropriate process for comprehension. “Information that is more abstract may prove to be more difficult to process and so divert cognitive resources from language processing. At the same time abstract information often implies a linguistic complexity that may further stretch the L2 reader’s resources (Weir et al., 2009; Green et al., 2010). It is therefore anticipated, that our analysis of the texts, in the Omani context, should reveal that the majority of language used should consist of more abstract rather than concrete words, as, in general, abstract words characterise academic texts. However, some caution is required here as the students in the foundation year in Oman are generally intended for more technical rather than academic courses and therefore there is less requirement for the recognition of devices that are more common in academic texts. Over-representation of abstract words in this context might lead to invalid interpretations of students’ scores and performance.

Based on these assumptions, Green and Hawkey’s (2011) study made the point that the passages used in IELTS tests should be more straightforward. However, it is important to make a distinction here. Abstract is not necessarily always equivalent to lack of straightforwardness. Some highly abstract words are straightforward for many students. For example, the word

‘energy’ is fairly high in abstraction but it is a relatively straightforward term to understand.

2.7.3.8 Content knowledge

The issue involved in content knowledge is the assessment of what influence the test taker’s background knowledge may have on the relative difficulty of a specific test task. There is some variation in the terminology used by different writers; for example, some writers approach this question of cognitive schemata in slightly different ways. This has resulted in wide and varied categorizations of schemata:

- Alderson (2000): Content schemata, knowledge of subject matter, knowledge of the world, cultural knowledge.
- Kirkland and Saunders (1991) demonstrated the significant role of cultural factors and religion (which involve both content schemata and affect) in reading comprehension.
- Grabe (1988): importance of knowledge of genre
- Grabe (1991): language background (Spanish, Asian, Arabic): “comprehension of texts may be culturally dependent according to the logical organization of the text” (Grabe, 1991, p.388).
- Chihara et al. (1989): background and culture as factors in EFL reading comprehension.

Thus, slightly different theories have developed to account for the influence of ‘background knowledge’ (Cohen, 1994; Alderson, 2000; Grabe, 2009). Some theories refer to *scripts* (interpersonal schemata), for common events e.g. eating in a restaurant, others refer to *frames* (ideational schemata in (Fulcher, 1998). Common to both is the conviction that the state of the reader’s knowledge influences process, product and recall (Alderson, 2000). Additionally, as previously mentioned, consideration needs to be given to the role played by the test taker’s characteristics in any assessment. Also, the role of test taker’s or reader’s knowledge and its relationship to the content of the text is well documented in the literature (e.g. Kirkland and Saunders, 1991; Clapham 1996; Khalifa 1997; Urquhart and Weir, 1998; Alderson,

2000).

A number of studies have provided empirical evidence that background knowledge effect in comprehension is very strong (e.g. Bransford et al., 1984 cited in Alderson, 2000, p.8; Klusewitz and Lorch, 2000). Fulcher's (1999) contention that the evidence does not well support the case for specificity re-echoes the earlier remarks of Nuttall (1996) and thus subject specific modules have been largely abandoned. Weir (2005) distinguishes between two types of content knowledge; the first is subject specific modules which have been shown to be, only in a very limited way, related to text difficulty and then, content familiarity, which again was found not to have a significant effect on text difficulty or test scores (Tan, 1990; Clapham, 1996). However, Khalifa (1997) came to an opposite finding that familiarity with the topic was a good predictor of difficulty (Weir, et al., 2009), and found support for this view in the earlier work of Alderson (2000). The issue is one of validity. If the view is taken that tests of L2 English should only be reflected in language couched in terms drawn from their subject specialism then there is not a perceptible problem of validity as, theoretically, no student has an advantage over another based on the texts. However, developments in testing have moved away from English for Specific Purposes (ESP) towards a more generalist approach and consequently, texts can be drawn from a wide range of content. Thus, a student who had a passion for astronomy might be seen to have a lucky but unfair advantage if one of the text contents was astronomical in nature. So there could be a question of validity based on test taker's content knowledge. However, that lucky advantage would be really confounded if all the texts were couched in astronomical language. This is confirmed by the research of Clapham who concluded, "…once the modules contained only 'specific' passages, background knowledge became proportionately more important. It might be hypothesised that if all the subtests had been 'highly specific', background knowledge might have made an equal or greater contribution to comprehension than language ability" (Clapham, 1996, p.205). The solution then is to have a wide variety of texts drawn from many academic subject areas (Enright et al., 2000). This has implications for testing in the Omani context in that a wider variety of texts

drawn from many academic subject areas needs to be included in the reading tests. On this very point, Cushing-Weigle (2000) was able to comment favourably on the MELAB proficiency test for academic purposes that “In terms of test content, the test writers have attempted to include content on a wide variety of subjects that would appeal to many different kinds of examinee, thus minimizing the risk that some examinees would be advantaged or disadvantaged by unequal content knowledge” (Cushing-Weigle, 2000, p.453).

What has been discussed previously about the various schemata types implies that, among the variety of texts, certain schemata types would be relevant in Oman such as cultural and religious factors. Finally, the comments made by Grabe (1991) about the effects of the L1 linguistic background might not be as important for these students as many of them are not yet proficient in their first language.

2.8 Summary

This study is more than a simple replication of the previous studies (Weir et al., 2009 and Green et al., 2010) as it includes variables for measuring certain task setting which were beyond the scope of those earlier studies. Actually, those previous studies drew attention to the need for the inclusion of these additional variables in future studies (see Section 2.5).

Whereas the previous studies focused on the assessment of L2 reading within a context where English was first language, this study considers the implications for the assessment of L2 reading within a context where English is not the first language. The importance of considering the learning context for English as a second language was highlighted by Oller and Tullius (1972) in a comparative study where one L2 group, who learned English in an ESL environment was tested against a second group who had learned English in an EFL environment. They found that the ESL group had a mean reading rate of 240 w.p.m. whereas the EFL group had a mean of only 182 w.p.m., a mean difference which proved to be significant ($p < .05$). Further support for the importance of the learning context is to be found in Favreau and

Segalowitz's study (1982), where L2 learners were found to be capable of reaching native speaker fluency. However, this study was in a bilingual context where both the L1 and L2 were widely spoken. Furthermore, the focus was on reading speed rather than comprehension. The setting of this current research is quite different in that English is being acquired as a foreign language where even the writing system is very different between L1 and L2.

In the light of the many models presented in this review, there emerges an understanding of reading as a complex activity that is multi-componential in nature. Different models were compared and contrasted, and the variables that affected the nature of reading were identified. The case was then made for adopting the model of Khalifa and Weir (2009) as the basis for researching the context validity of reading assessments in the foundation programmes in the Omani Colleges of Technology. Utilising the variables earlier identified in Weir's (2005) test validation framework, context validity is evaluated in this study using scalar measures or expert judges in the case of those variables which were not directly amenable to scalar measurement. The distinct nature of these variables in a second language context was tested by correlations and factor analysis in (Chapter 4).

Scalar measurement and opinions of expert judges, already utilised to great effect in Green et al. (2010) and Weir et al. (2009), are extended by the inclusion of a sample of students whose views were elicited during a natural experiment utilising Verbal Protocol Analysis (VPA).

In the light of the insights gained through the various context validity features, it was expected that a fair evaluation could be made of the extent to which the test tasks at level four on the foundation programme were aligned to the reading texts encountered during the first year of academic study.

Chapter 3 Research design and methodology

3.1 Introduction

This chapter presents the research design for the study which is focused on determining the extent to which the current reading test tasks in Oman were representative of texts encountered by students at first year academic level. In order to assess the current test tasks, a number of methods were employed.

First, a comparison and analysis was conducted to determine how well the test tasks in reading and comprehension in the Level Exit Exam (LEE) were aligned to the reading texts encountered at First Year Academic (FYA) level. The particular method utilised for this purpose was an automated text analysis. A number of variables was not directly amenable to assessment using this method and, for these, a panel of expert judges were asked to draw on their expertise in deciding the issue.

Second, as it was desirable to augment the findings from the automated analysis and expert judges' considered opinions, students were also invited to participate in a natural experiment, using a Verbal Protocol Analysis (VPA). As students are the main end users of the test it was expected that a valuable contribution could be made to the research by investigating the cognitive processes they employed in addressing the test task.

This chapter presents the research question and sub-questions. How the research has been designed is then discussed followed by a description of the instruments and a justification for their use in this study. The twin principles of validity and reliability are then considered and ethical issues are identified and addressed. Finally, the key points of the research design and methodology are summarised.

3.2 Research question and sub-questions

Although Cohen et al. (2011) point out that it is not always necessary for a piece of research to be driven by a research question (e.g. some qualitative studies), nevertheless, a central research question is helpful in giving focus to the overall study. However, it is important that the central research question should be open-ended using words such as 'how' or 'what' *'to convey open and emerging design'* (Creswell, 2009, p.130). The research also should have a focus and with a special context in mind (Robson, 2011); in this study the context and focus is the Foundation Programme in the CTs in Oman. Robson (2011) points out how a question first arises but then needs to be anchored in a particular situation and set of circumstances stimulated by theoretical concerns (Robson, 2011). This process resulted in the formulation of the following research question:

"How closely do the texts in the reading section of the Level Exit Exam (LEE) at Foundation Program level resemble the reading texts that students encounter at first year academic (FYA) level in Oman?"

Many authors recommend further articulation of the research question through sub-questions. In particular, Robson's advice was followed by ensuring that sub-questions:

- "are clear and unambiguous;
- show the purpose(s) of your project (to explore, describe, explain and/or empower);
- are answerable – and point to the type of data needed to provide answers;
- are not trivial; and
- form a coherent interconnected set (they are not an apparently random collection)" (Robson, 2011, p.62).

Drawing on the literature review, the following three research sub-questions (RSQ) were proposed:

1. What are the cognitive processes by which students engage with the texts and tasks in the reading tests?
2. How closely do the reading texts in the Level Exit Exam (LEE) of the FP reflect those encountered at first year academic (FYA) level texts?
3. How closely do the reading tasks in the LEE of the FP reflect those encountered at FYA level?

The literature review revealed the variables by which task setting and linguistic demands could be gauged (see Chapter 2 Section 2.7). For test task setting these were:

- Response method
- Weighting
- Knowledge of criteria
- Order of items
- Channel of presentation
- Text length
- Time constraints

The variables identified in the literature for linguistic demands were:

- Overall text purpose
- Writer-reader relationship
- Discourse mode
- Functional resources
- Grammatical resources
- Lexical resources
- Nature of information
- Content knowledge

3.3 Research design

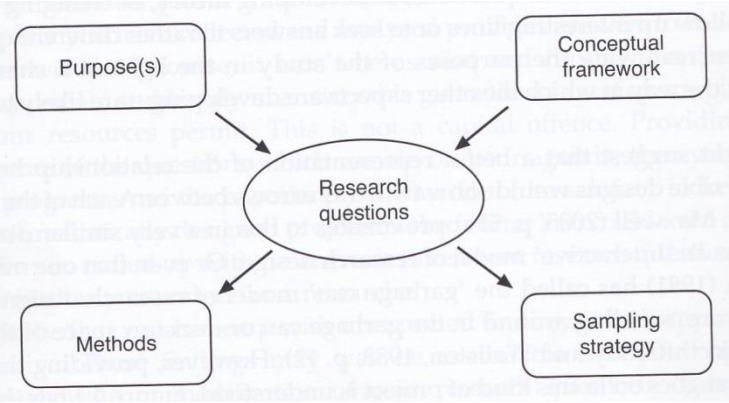
The research questions determine the design and the scope of the research as a whole (Gilbert, 2008; Green, 2008). In the literature on research design, there is a broad range of designs which have emerged over time and these are generally categorized as quantitative, qualitative or mixed methods (Creswell, 2003, 2009). Another approach to categorising research designs

is that of ‘fixed’ or ‘flexible’ and ‘multi-strategy’ design approaches (Robson, 2011). Miller and Salkind (2002) present research designs categorized into seven basic types. Cohen et al. (2011) consider research design from the perspective of the underlying paradigms on which they are based. This study is based on a multi-strategy method design following (Robson, 2011), which was based on the earlier works of (Creswell, 2003; 2009). Although Cohen et al. (2011) consider the advantages of multi-strategy approaches, these tend to be less flexible in categorisation. For this reason, Robson’s design is preferred in this study because, as well as being clear and easy to follow, it is also a much more flexible and adaptable one for the purposes of this study.

The value of multiple methods in second language testing research is also becoming increasingly widely recognised (e.g. Clapham, 1996; Bachman, 2000). A multiple strategy design was considered appropriate for this study for two reasons; firstly, a single method could not be found to embrace all of the variables that required testing and, secondly, a single method would not have fully answered the research questions. Consequently, the design choice embraced an automated text analysis for a comparison to be drawn based on test tasks compared with academic texts. The considered judgments of experts in this field were also be employed, as well as capturing, through self-reporting, the mental processes which students engaged in as they tackled the test tasks. The use of both quantitative and qualitative methods was used to address the research questions and, combining these two methods, promised to provide more comprehensive answers than would have been possible using a single method approach.

The framework in Figure (3.1) which was presented in Robson (2011) was found to be helpful in arriving at a suitable framework for the research design:

Figure 3.1 Framework for research design



Adopted from Robson (2011, p.71)

- The *purpose* of this study was essentially evaluative in orientation in that current practice of second language testing was being investigated and assessed in order to collect empirical evidence to address the research problem and aims of the study (see Chapter 1 Sections 1.2, 1.3 and 1.4).
- The *conceptual framework* informing the research design was pragmatic in nature and based on the assumption that valid sources of knowledge need not be strictly compartmentalised but that qualitative type data (expert judges' opinions) could amplify other types of data i.e. data obtained by VPA and automated software analysis. However, the study is fundamentally rooted in assumptions of post-positivisim in which objective truth or approximate truth is seen to be attainable (Creswell, 2009). In other words, in this study, truth is assumed to be mind-independent reflected in the use of descriptive and inferential statistics in order to answer the central research question i.e. how closely the LEE tests in FP resembled the FYA texts. Accordingly, the study is largely deductive in nature in that it is driven by the theoretical model of Khalifa and Weir (2009), a model which is tested and validated in this study in the context of L2 learners in a L2 setting. However, a pragmatic solution is adopted whereby other sources of knowledge outside of a post-positivist paradigm are also seen as valid yet different sources of data. Variables which were not amenable to direct measurement quantitatively (e.g. writer-reader

relationship) were measured through a qualitative method. However, the analysis of this data was quantitative in nature.

- The *method* was predominantly quantitative and although it did not have an explicitly stated hypothesis, the aim was to address the paucity of empirical evidence relating to the Omani tests using the theoretical model of Khalifa and Weir (2009). The design was flexible rather than fixed (Robson, 2011) in that the study did not rely on a single method of collecting evidence.
- The *sampling strategy* entailed a selection of reading test tasks from recent LEE papers and a sample of reading texts selected from the main textbooks in the FYA main academic faculties (IT, business studies, and engineering). The second phase involved the judgements of experts in the field of language testing and other relevant areas. The third phase entailed a sample of students from two of the colleges participating in a natural experiment designed to capture their thought processes as they engaged with the test tasks. Details of the sampling strategy for both methods receive a more thorough treatment in Section 3.4.2 and Section 3.5.2.

3.4 Methods and instruments 1: A natural experiment using Verbal Protocol Analysis (VPA)

In this section, the method and instrument of verbal protocol analysis (VPA) are presented and justified. Aspects of design and content, sampling strategies and limitations are discussed.

The findings of the two phases of data collection (text automated analysis and expert judges) were expected to shed light on the nature of the test tasks and academic texts and their relevant contextual features. The socio-cognitive model of reading emphasized the importance of the test taker in interacting with the test tasks. The processes which the test taker employs could not be examined in the first phase. Earlier models tended to focus on the text only to the neglect of the test taker and this limitation has been highlighted in the critique in the literature review. For this reason it was felt necessary to include a phase in the research which focused on the role of

the test taker and the cognitive processing which was involved in engaging with the test tasks. It was anticipated that this process-based phase would involve the students, the main focus of any assessment, and would thus enhance the findings of previous studies which were confined to analyses of the texts.

The use of reports based on VPA has grown in popularity mainly due to its method based on a process-oriented approach (Embretson, 1983; Messick, 1995; Ross, 1997) and deemed to be valuable in second language research (e.g. Seliger and Shohamy, 1989; Cohen 1994; Cohen, 1998; Brown and Rodgers, 2002; Taylor, 2004). VPA is very useful in L2 testing for investigating test taking processes and have been considered of great importance in drawing inferences related to the various cognitive processes involved in test taking (Green, 1998). These are defined as "test-taking processes that the respondents have selected and that they are conscious of, at least to some degree" (Cohen, 2000, p.129). Examples of various studies investigating the test takers' processes in the testing of reading and comprehension include (e.g. Norris, 1990; Sasaki, 2000; Yamashita, 2003; Leow and Morgan-Short, 2004; Rupp et al., 2006; Cohen and Upton, 2014). This is all the more important in terms of this study which adopts a cognitive-based model. Seliger and Shohamy (1989) report a number of studies which used VPA in second language acquisition:

- Brown (1983) collected verbal reports on how experts formulate written summaries of reading texts
- Mann (1983) elicited subjects' self-reports
- Olshavsky (1977) investigated reading strategies by obtaining verbal reports from readers

VPA was also used as a retrospective verbal report protocol with non-native English students (Cohen and Olshtain, 1993). A retrospective VPA was considered to be the best approach in the context of this study. However desirable it might be, conducting VPA introspectively in the real life situation of an actual examination was considered to be impractical for a number of reasons. First of all, it would have been highly intrusive into an exam which was of high stake significance for the learners. Having to report their mental

processes in responding to test tasks could have been captured by sound recordings but would have been a distraction from the test itself both for the test taker and for other students. Capturing their cognitive processes immediately after the exam was another possibility but, again, there were difficulties because the students would have just completed a three hour exam and it would have been difficult to have expected them to revisit parts of the exam and recall their thinking. Secondly, such recall might have been confounded by other variables related to listening and writing exams because they are usually taken at the same time (see Chapter 1 Section 1.1). Nor would it have been helpful to ask them to recall their thoughts some hours or a day or two later as the immediacy of the situation would have been lost and, once again, other variables would have constituted interference in the experiment (Seliger and Shohamy, 1995). For these reasons, it was decided to have a simulated exam sometime later.

It is readily admitted that a simulation is not the real thing and students might not have been well motivated due to the fact that this simulated exam would have had no consequences in terms of their progression. However, there were a number of advantages to conducting a natural experiment in the form of a simulated test. Firstly, the simulation was based on reading skill test only, thus minimising the possibility of irrelevant or confounding variables. Secondly, the problem of interruption would not arise as would be the case in a real test situation. Provided the researcher could succeed in motivating the participants to approximate as much as possible to a real life situation, it was considered to be possible to capture the 'thinking aloud' processes that the students engaged in when answering the questions. Motivation was achieved through briefing the participants prior to the simulation explaining the purpose of the exercise and its potential benefits for future generations of students. There could also be potential benefits for the participants themselves as it was possible that some recommendations from the analysis and findings might have implications for their first year academic reading (Cohen, 2000). Nevo (1989) cited in Cohen (2000, p.132), reports that L2 students who took part in self-reporting actually benefitted from the exercise by becoming aware of the various strategies adopted in addressing the

tasks.

The next issue concerned which instrument would be most appropriate for collecting data relating to the students' mental processes. A number of qualitative instruments were considered. First of all, interviews were considered but, in order to achieve an immediate recall, this would have meant a one to one correspondence of interview to participant which was impractical. It was also considered impractical to interview the students at say 15 minutes intervals as the immediacy of the recall would have been lost for most students. In fact, Ericsson (2006) has argued cogently that the recall should be within seconds of completing the task. Interviews would have been inappropriate for answering the research question which should not have been based on participants' interpretations of how tasks were addressed but rather the simple 'thinking aloud' process that accompanied their answering of the questions. Green (1998) argues against such in-depth probing as might happen in interviews: "Most important is that protocol analysis requires subjects to express their thoughts, but not to infer the processes that produced those thoughts" (Green, 1998, p.4). It was simply the recall of the thought underlying the performance of the task that was sought but not a deeper interpretation of that. Of course there was some inferential analysis later, but this was on the part of the researcher, not of the student (Green, 1998). Another method which was considered was that of observation. However, this was not considered appropriate on the grounds that interpreting the subtleties of facial expressions and body language would require an expertise beyond that of the researcher. Besides, while some of the variables might be amenable to being observed (e.g. coping with time constraint), other variables would not have been so easily amenable to such interpretation (e. g. to what extent posture, movement facial, expressions…etc. could reveal meanings attached to cultural knowledge, for example, was very much open to question). Another possible approach would have been conducting a phenomenological enquiry to capture the essence of, for example, the experience of taking a reading test. Again, although it does have a mechanism for eliminating researcher bias (i.e. the *epoche* method, Moustakas, 1994), it was still seen as highly interpretative in terms of the meanings attached by the participants

to the phenomenon, which was not the desired outcome for the research sub-questions.

It was finally decided to opt for a questionnaire with questions based on the variables identified in the literature review. This instrument was also adopted in a similar study by Weir et al. (2009), although the overall thrust of the questionnaire in this study was quite different in design and content. Statements were preferred to questions as, not only were participants' interpretations not required, they were to be avoided in the light of Green (1998) and the note of caution issued by Seliger and Shohamy: "the need for additional verbal processing may interfere with the processing that is being commented on" (Seliger and Shohamy, 1995, p.170). All that was required was a simple and immediate recall of how the participants thought their way through the various questions and tasks and this was best achieved by a questionnaire. It was an ideal solution for capturing data immediately following the test from a large group of participants so that each participant could think aloud and respond to the statements using immediate short-term memory recall. Another advantage of this method was that it was possible to conduct it in more than one location. Data from more than one location was expected to strengthen the validity of the research. Also, questionnaires were considered as allowing for intra- and inter-rater reliability (Seliger and Shohamy, 1995).

It can still be argued that a simulated test is not like the real test. However, the difficulties of capturing the required data from a real exam have already been highlighted and there is no known method of doing this in a way which does not interfere with performance in the exam itself.

3.4.1 The questionnaire: design and content

The questionnaire in this study was inspired by the verbal protocol method as presented in (Pressley and Afflerbach, 1995; Green, 1998). A questionnaire was designed which aimed at capturing the thought patterns of students as they attempted the various tasks in the reading LEE test in Oman. The questionnaire developed by Akmar Saidatul Zainal Abidin

presented in Weir (2005, pp.224-225) served as a starting point for the development of the instrument although their study was focused on the speaking test. By identifying variables related to reading it was hoped that an instrument could be developed that would make it possible to capture the thinking processes by which students answered the test questions. The ideal setting for a verbal protocol which relies on immediate short-term memory recall would be during the actual test itself. However, this was considered to be impractical as has been discussed earlier. It was therefore decided that a simulation would be the next best option with students taking the reading section of the English test in a natural experiment in which they reported how they thought out their answers.

The rationale for choosing a questionnaire was its usefulness for gauging thoughts and opinions across a relatively large sample (e.g. Nardi, 2003). Admittedly, this sample was not large enough to be considered a strictly representative one drawn from a population of over 11,000 students scattered in different geographical locations (see map in Appendix 1). There was a number of reasons why this was so. Being unable to collect data from an actual test made it necessary to conduct an experiment using a simulated test sometime in the weeks following the actual test. This necessitated reliance on students' willingness to participate on a voluntary basis. Furthermore, it was unlikely that a large number of students would have been available at that time. Finally, the verbal protocol needed to be conducted in controlled conditions so that students' thoughts could then have been collected by means of their responses to the statements in the questionnaire. Since this involved immediate recall, it was deemed to be necessary that the students were talked through the process and briefed about the purpose of the verbal protocol. This was considered to be best accomplished by on the spot instructions followed by administration of the questionnaire. Thus, an online questionnaire would not have been appropriate even though it would have had the advantage of covering a larger and more representative sample.

Guidelines of writing good surveys and producing good questionnaires were strictly followed as advised in (Brown, 2001). These were:

- Thinking about the form
- Thinking about the meaning
- Thinking about respondents
- Write good questions
- Order the questions rationally
- Format the questionnaire for clarity
- Write clear directions
- Edit carefully

Also, regarding the framing of the statements, there is an abundance of literature (e.g. Oppenheim, 1992; Gillham, 2000; Nardi, 2003; Procter, 2008; Simmonse, 2008; Wisker, 2008). They commonly emphasize: consistency in the framing of statements, avoidance of technical language, avoidance of qualifying adverbs which could confuse the respondents' choice of rating and the directionality of the statements. The past tense was used consistently as data was being gathered retrospectively through short-term memory recall. The questions adopted a positive stance avoiding negativity "because the presence of a negative or double negative in a question can lead to misinterpretation and confusion" (Brown, 2001, p.47). Also mentioned was the avoidance of double-barrelled questions; for example, statements 9, 10, 11, and 12 had originally been one composite statement. It was realised, however, that the thinking involved for each individual variable could be different and a composite statement would thus be confounding for the respondents. Care was also taken to avoid leading or loaded statements. For example, statement (24) originally read '*Recognising that this was an informative passage helped my understanding*' but it was realised that this involved leading the respondents regarding the purpose of the passage being used instead of leaving the test taker to make this decision.

Having taken all these considerations into account, there was confidence that the instrument would meet the requirements for validity, reliability and the elimination of bias.

The guidelines for data preparation and collection outlined in Green (1998) were followed in preparing statements which would reflect students' thought processes in responding to variables which measured reading. The reading section of a past test paper was selected (entitled Zanzibar from 2013 tests, see Appendix 2). Following Green (1988) the process of task identification and analysis was followed. Before a draft was developed, hunches were first arrived at through brainstorming based on the literature review of the variables (Brown, 2001). These were set out in order and then were critiqued and refined from the point of view of their suitability for VPA. Next, the task and items in the test paper were analysed and the particular types of reading and levels were identified according to the matrix of reading in Khalifa and Weir (2009). Doing so enabled the matching of sets of statements which might capture the thought processes involved in particular types of reading at various levels.

This resulted in a set of statements relating to the particular reading test tasks and these are presented in Table (3.1) which contains a column to report the particular variables being tested:

Table 3.1 Initial statements and variables

sn	Statements	Intended variables being tested
1.	I found the test instruction easy to understand	Rubric
2.	I agree with the number of marks given for each task	Weighting
3.	I was quickly able to find the information required to answer the questions in PART A.	Scanning local expeditious If agree they were reading expeditiously if disagree they were reading carefully which is wrong
4.	It was difficult to decide whether to skim (fast read) or read carefully the whole passage in order to answer questions in (PART B).	Examining discernment Skimming expeditious reading
5.	I had to read other sentences carefully in addition to the sentence on line 16 in order to answer question 8.	Examining careful reading at global level If they agreed then they are utilising careful reading if disagree then they have not used careful reading which was required Also examines Grammatical resources – syntax the effect of the coma after adjectives for example how the comma after prosperous changes the meaning of the sentence line 17/how the comma after prosperous is necessary to convey the exact meaning intended
6.	I was able to answer question 9 by carefully reading the sentences in line 17 and 18.	Careful reading at local level. If agree they have used careful reading but it is not possible due to the sentence that is referred to does not contain information so test takers must have guessed This also examines (Grammatical resources)
7.	I was unsure whether long detailed answers or short answers were required in answering (PART D).	Examining rubric

8.	I had to quickly search the whole passage to find the information required to answer question 10 (PART D).	Examining expeditious local level search If agree they did they have correctly applied expeditious local.
9.	I had to quickly search the whole passage to find the information required to answer question 11 (PART D).	
10.	I had to quickly search the whole passage to find the information required to answer question 12 (PART D).	
11.	I had to quickly search the whole passage to find the information required to answer question 13 (PART D).	
12.	To answer questions 18, 19, 20, 21, and 22 in (PART F) I had to read and comprehend the whole passage.	Deletion gap filling Test takers do not need to read the whole passage to be competent readers
13.	I found the variety of question types (e.g. true/false, short answer, gap filling, etc.) helpful in allowing me to show my skills.	Response methods
14.	Question 8 should have been awarded more than 1 mark	Weighting
15.	It was helpful to know how many marks allocated to each item.	Knowledge of criteria
16.	The knowledge of the number of marks given to each item affected my time planning and execution	Knowledge of criteria
17.	Before starting to answer the questions it was important to plan how much time should be spent on each item.	Knowledge of criteria
18.	It was important to answer the questions in the order they were presented.	Order of items In this passage no
19.	A time line of main events accompanying the passage would have aided and supported my passage comprehension.	Channel of presentation
20.	It was not necessary to read every word of the passage in order to understand its meaning.	Passage length Skimming and scanning
21.	The length of the passage made it difficult to read and understand in the time provided.	Passage length

22.	I looked at the questions first before deciding whether to read the passage carefully or quickly.	Discernment/strategy
23.	40 minutes was sufficient time to answer all the questions.	Time constraint
24.	Recognising the overall purpose of the passage helped my understanding.	Overall passage purpose Main purpose
25.	This article is probably taken from a secondary school history passage book.	Writer-reader relationship No
26.	This passage is suitable for Omani students.	Writer-reader relationship
27.	The organisation of the five paragraphs in the passage helped my understanding and comprehension.	Discourse mode (cohesive devices) The passage is well laid out logically and chronologically coherent
28.	The passage presents the historical development of Zanzibar in an easy way to understand.	Discourse mode Historical genre
29.	The use of 'At the moment', 'but' in lines 33 and 34 helped me to understand the present tourist situation in Zanzibar and future projections.	Discourse mode Rhetorical device If it helps that a rhetorical that worked
30.	In lines 9 and 10, the structure 'though….', 'now' helped me to understand current developments of conservation of trees.	Discourse mode
31.	In line 7 the sentence 'the climate is tropical and humid', required the sentence that followed it for a more in depth understanding.	Discourse mode Rhetorical device
32.	I understood from question 12 that the Sultan moved the entire city of Muscat brick by brick to Zanzibar.	Functional resources If disagree then test takers in Oman likely have no issue in dealing with functional resources (Ideational function) Description
33.	The 'colourful history' of Zanzibar line 27 is due to the amount of sunshine the island receives.	Functional resources If agree they are wrong (Imaginative function)

34.	The sentence on line 30 would be more meaningful if it read 'the people in Zanzibar also manufacture clove oil and woven goods'	Functional resources If agree they cannot cope with such type of language (Imaginative function)
35.	It was easy to understand what was referred by 'it' in line 7	Grammatical resources Pronoun - able to understand
36.	Although paragraph 1 had only two short sentences it was difficult to decide which of the three options best expressed the main idea.	Grammatical resources Examining whether short sentences can still be difficult to comprehend
37.	Paragraph 3 had long sentences which made it difficult to decide what the main idea was (question 6).	Grammatical resources
38.	Because I already knew the meaning of 'extensively' in question 8 I did not need to refer to the passage.	Lexical resources
39.	It was necessary to refer to the passage to work out the meaning of prosperous in line 17.	Lexical resources
40.	My previous knowledge of the meaning of most of the words in the box in (PART F) meant that I was able to answer this question without referring to the passage.	Lexical resources
41.	In (PART F) it was not necessary to understand the exact meaning of words such as 'culture', 'environment', 'tropical' and 'humid' in order to complete the paragraph.	Lexical resources
42.	Most of the words on the passages were concrete, meaning factual straight words and easy to understand, e.g. car, apple rather than abstract words e.g. economy	Nature of information Abstract or concrete
43.	The article was mostly familiar to me due to my history of the connections between Zanzibar and Oman.	Content Knowledge
44.	From my geographical knowledge I already understood the role of 'coral reef' in question 10.	Content Knowledge General background knowledge
45.	Even though there were no headings, it was still easy to understand the passage.	Channel of presentation

In the questionnaire, it was of utmost importance that statements were articulated which reflected the issues highlighted by the research question and its associated sub-questions.

Of the seven questionnaire types outlined by Youngman (1982, 1994) cited in Bell (2010, pp.211-212), a rating scale was considered to be the most appropriate for measuring the strength of agreement or disagreement with the statements as advised in Nardi (2003) and Simmonse (2008).

The first page of the instrument (see Questionnaire in Appendix 3) introduced the respondents to the aim and purpose of the questionnaire. Next, some ethical issues followed: assurance of confidentiality, the voluntary nature of participation, and the right to withdraw from the study within two weeks of data collection. An assurance of anonymity was given, as well as an explanation of the structure of the questionnaire and contact details of the researcher. Finally, the respondents were thanked for their cooperation.

The second page (Part I), collected some factual information from respondents. The first three questions were about their college, gender, and subject specialisations with a view to making possible comparisons and generalisations at analysis stage.

Part II, the questionnaire proper, consisted of a set of statements to which the participant responded by selecting from a *Likert* four-point scale, ranging from strongly agree to strongly disagree. The choice of a four-point scale avoids a 'sitting on the fence' position. Note was taken of the criticism made against this approach by writers such as (Oppenheim, 1992; Procter, 2008) who pointed out that the neutral option might sometimes genuinely express the true position of the respondents. However, the neutral option was not appropriate in this situation as the students had already completed a task and would have accomplished that task one way or another. Being neutral was clearly not an option. It is admitted that a particular respondent might have been unsure in addressing a particular statement but the likelihood was considered to be small as the recall was occurring immediately after completing the simulated test. This was checked in the piloting and it was

found to work (this is dealt with more fully in Section 3.6.1 under pilot study for VPA). The point is that attitude was not being measured here so it was most unlikely that there would have been ambivalent responses. What was being captured was the thought process underlying a completed task.

The statements were developed by refinements of elements based on the literature review and represented the end result of the task identification and analysis of the tasks as set out in Green (1998). These statements were then arranged in such a way as to produce a logically coherent structure. The challenge here was to present certain key ideas with very fine meanings into statements presented in simple language. For example, statement (1) was about the 'rubric' of the tasks, but the use of this technical word was avoided.

3.4.2 Sampling

The sampling strategy was based on students volunteering to take part in an experiment. The aim was to have as large a sample as possible as recommended by many authors (e.g. Nardi, 2003). This is also well supported in educational research as "small effects can reveal themselves which might otherwise be lost with small samples, even though the trade-off here is that, with large samples, it is easier to achieve statistical significance" (Cohen et al., 2011, p.325). It was considered impractical to attempt to stratify by gender nor was this considered to be essential as long as a reasonable ratio of male and female students were included which broadly reflected the gender ratio on the campus. The sample was drawn from two of the 7 Colleges of Technology. The Higher College of Technology in Muscat represented a metropolitan setting while Ibra College of Technology represented a more rural environment.

3.4.3 Limitations

It is admitted that the sample was not strictly representative. However, the experimental nature of the method has already been discussed and the impracticality of stratification across gender and geographical locations were

discussed. Nevertheless, it is not considered that validity is in question as the focus of the experiment was to capture thought processes involved in completing test tasks. The possibility of gender differences and differences based on subject specialisations were catered for by ensuring a reasonable mix of male and female, and cross college specialisations of students.

The possibility of a low proportion of students volunteering was considered and addressed by seeking the cooperation of department heads and teachers in motivating the students to participate. Secondly, another potential disadvantage was that some locations or groups might be over-researched and this could have resulted in questionnaire fatigue leading to a low return rate. Thirdly, although there generally existed a risk of a low return rate due to questionnaire fatigue among over-researched groups, this was not considered to be a serious issue in this research as the Colleges of Technology had not been the subject of significant amounts of research in recent years.

Lack of control of the environment and of the order of the questions are often cited as possible limitations (Brown, 2001). However, this was a controlled experiment and those limitations were not thought to have been significant. Other possible disadvantages mentioned in the literature included artificiality, rigidity and the impersonal nature of the questionnaire (Brown, 2001). These were not significant issues as the researcher was present and it was conducted in a personal manner with the respondents being assured that their candid responses would be greatly valued and of assistance for future testing of reading. The questionnaire statements, although being rated on a *Likert* scale, were not in themselves rigid but were actually the result of a nuanced approach to understanding the underlying thought processes based on models of reading from the literature review.

3.5 Methods and Instruments 2: Automated text analysis and expert judges

In this section, the methods and instruments of automated text analysis and expert judges used in this research are presented and justified. Aspects of

design and content, sampling strategies and limitations are also discussed as well as an explanation of what is being measured by each instrument.

Three different computational tools were used to analyse a sample of test tasks for comparative purposes with a representative sample of academic texts which were drawn from first year program text books. The analysis provided the basis for measuring the contextual features obtained from the model of Khalifa and Weir (2009) and comparisons were made between the scores obtained for test tasks and the scores obtained for academic texts in order to address the research question relating to how closely the test tasks resembled the academic texts on each of the variables.

Certain variables were not directly amenable to measurement by means of automated text analysis. For these variables, expert judges were asked to offer their considered opinions on the variables in both sets.

The three instruments which were used for text analysis were:

1. Coh-Metrix (Graesser et al., 2004, McNamara et al., 2005; McNamara et al., 2012)

2. Web VocabProfiler (Cobb, 2003)

3. WordSmith Tools (Scott, 2006)

The Coh-Metrix instrument measures various grammatical resources and the general cohesion of a text. It has a number of advantages over previous tools. Firstly, it is capable of analysis based on one single measure rather than an array of measures. Secondly, it is more up to date in meeting the needs for understanding discourse and text including text difficulty and readability. Thirdly, it has been found to measure cohesion validly and reliability.

The instrument itself has been well validated and a number of studies of this validation are presented in (McNamara & Graesser, 2012):

- McCarthy et al. (n.d.)
- McNamara et al. (2010) established the validity of Coh-Metrix as a measure of cohesion differences in texts

- A previous study, Duran et al. (2007) validated Coh-Metrix based on high school science, history and narrative texts (Duran et al., 2007, p.193). This study also found that Coh-Metrix produced highly predictive indices of text genre. This is important for the current research as all of these variables were to be tested.

In second language assessment, the two main works (Weir et al., 2009; Green et al., 2010) which principally informed this study have also analysed test tasks and undergraduate university texts using Coh-Metrix.

In comparison to other measures such as Flesch-Kincaid and Lexile scores (relating text characteristics to the performance of readers on cloze tasks), which focus much on words and sentences (i.e. number of syllables/words and/or word/sentences), Coh-Metrix takes a further step by having an overarching goal of looking into the link between words/sentences and sentences/paragraphs. Additionally, it focuses on cohesion such as ideas and topics as well as the use of connectives (McNamara and Graesser, 2012). It not only measures lexical diversity and connectives but also goes deeper into the analysis of syntactic complexity, by noting the number of modifiers in noun-phrases and the number of words that occur before the main verb in a sentence (McNamara and Graesser, 2012).

The second instrument is Web VocabProfiler developed by Cobb (2003) through rigorous consultations with a wide sample of different stakeholders including more than 1500 learners, teachers and researchers worldwide (Cobb, 2013). It is a computerised program that carries out lexical analysis. Any text can be loaded into it and the words within the text are divided into four categories in terms of frequency: (1) the 1000 most frequent words, (2) the second thousand most frequent words, (3) English academic words, and (4) whatever remains as words which do not fit into any of the other lists. Thus, it is a measure of the proportions of low and high frequency vocabulary which would be used by a native speaker. Typically, a native speaker would have a profile in which 70% would be drawn from the first thousand, 10% from the second thousand, and 10% from academic words (Cobb, 2013). It has been employed in various studies for different purposes and is acknowledged as software which is reliable having undergone a validation

process as a research instrument (Laufer and Nation, 1995). Additionally, a number of different studies used Web VocabProfiler in texts analysis (Meara, 1993; Meara et al., 1997; Meara and Fitzpatrick, 2000; Cobb and Horst, 2001) (Cobb, 2013).

A major use of Web VocabProfiler is in evaluating reading texts' suitability and how they align to different learners' levels at Academic Word List Level (AWL) (Cobb, 2013).

Finally, a third computational tool used in this study is WordSmith, which provides multi-task tools:

- WordList: provides a list of the words or word clusters in a passage and can present these in alphabetical/ frequency order.
- Concord: this sets individual words in the context in which it occurs (proximately to certain other words)
- Identifies the keywords in a passage

These tools have been used widely across the world by different users ranging from students to teachers and include such eminent institutions as Oxford University Press for dictionary preparation using lexicographic work (Scott, 2014). Similar to the Green et al. (2010) study and also this study, WordSmith can be used to evaluate the Type/Token Ratio (TTR). TTR measures the total number of words in a passage and then calculates the number of repeats and this gives the number of different words in the passage which will be less than the overall word count. The total number of words is called a token and the total number of different words is referred to as types. Thus, if a passage has 10,000 words, it is considered to have 10,000 tokens but when repeat words are taken into account there might be only 3,500 different words in the passage (Scott, 2006). Thus, there would be 3,500 types. The ratio of types to tokens would then be 35%. Green et al. (2010) points out the importance of having similar sized passages in terms of word count in order to make valid comparisons. In this research, the passages from exams and from textbooks were approximately 500 words in length. This made it possible to make valid comparisons based on TTR.

Green et al. (2010) also points out that, generally, the higher the TTR the more demanding a passage is to comprehend.

A wide range of publications ranging from journal articles to books have been written about and have used WordSmith (Scott, 2006).

This phase of the data collection followed closely the method outlined in Green et al. (2010) and Weir et al. (2009) although the range of variables was extended here to include 'task setting'. In this phase, the focus was on the content analysis of the LEE test tasks as well texts of undergraduate level (FYA). Although content analysis is perceived to be highly rooted in quantitative-based approaches (Bryman, 2001), in this part of the study, the data was reliant on judges' opinions, which to some extent reflected qualitative judgments. Nevertheless, the analysis was based on quantitative approaches as a prepared checklist was used for the judges to indicate their decisions (see Appendix 8 for checklist).

The use of expert judges in L2 research is widely acknowledged as a useful method (e.g. Bachman et al., 1996; Bachman 2004; Davies, 2011) as well as for testing and content validity (e.g. Pilliner, 1968; Davies, 1990; Green et al., 2013).

3.5.1 Automated analysis software and judges' checklist: design and content

In this sub-section, the variables that were to be tested were identified following (Green et al., 2010). Also, a checklist based on Weir et al (2009), and Weir (2005, pp.56-84) was designed for the use of the expert judges in making their evaluations.

It is useful, at this point, to list all of the context validity features that were to be examined and to indicate in exactly what manner they would be tested in this research. Table (3.2) below sets out these features and indicates by which method they were to be tested (judges, software or both):

Table 3.2 Contextual features and methods of testing

	CONTEXT VALIDITY FEATURES	METHODS	
		Expert judges	Automated analysis software
Linguistic demands: Task input and output	▪ Overall text purpose	√	
	▪ Writer-reader relationship	√	
	▪ Discourse mode	√	
	▪ Functional resources	√	
	▪ Grammatical resources	√	√
	▪ Lexical resources		√
	▪ Nature of information	√	√
	▪ Content knowledge	√	
Task setting	▪ Response method	√	
	▪ Weighting	√	
	▪ Knowledge of criteria	√	
	▪ Order of items	√	
	▪ Channel of presentation	√	
	▪ Text length	√	
	▪ Time constraints	√	

The table below contains a column for the variables that were to be tested by expert judges' opinions and a column for variables that were to be tested by the different software discussed above (see Section 3.5). For the variables examined by the judges, statements were constructed to enable the judges to express their judgements. These statements are presented in Table (3.3):

TABLE 3.3 Variables were examined by expert judges and different automated software

Variables for judges	Variables for software
Linguistic demands **Task input and output**	
1. Overall text purpose √ The category that best describes the overall text purpose	**N/A**
2. Writer-reader relationship √ Identify the intended audience/reader of the text that is targeted by the writer. Hyland's 2005 cited in Weir et al. (2009, p.138 and p.110)	**N/A**
3 Discourse mode √ Genre: Identify the most appropriate category for the text. Whether it is textbook, magazine/newspaper article, research/academic journal article, report Weir et al. (2009, p. 137) √ Rhetorical task: Identify the most appropriate category for the text. Exposition, argumentation/persuasion/evaluation, historical biographical/autobiographical narrative √ Pattern of exposition: Identify the pattern(s) used in the text.	**N/A**

Define, describe, elaborate, illustrate, compare/contrast, classify, cause/effect, problem/solution, justify √ Rhetorical organisation: The organisational structure of the text is… explicit or not explicit	
4 Functional resources √ Identify the most appropriate category for the text. Ideational, manipulative, heuristic, imaginative (Bachman and Palmer, 2010; Weir, 2005)	**N/A**
5 Grammatical resources √ Grammar: The sentences in the text are: range from mainly simple sentences to mostly complex sentences Cohesion Throughout the text, are relations between the ideas explicitly marked through reference, conjunctions and connectors or are such relations not explicit? Whether explicit or not explicit (Weir et al., 2009, p.137)	√ **adopted from Green et al. (2010)** **Grammatical complexity** G1: average number of words/sentences G2: average number of sentences/paragraph G3: the proportion of words included in noun phrases G4: number of modifiers per noun phrase (concerns the occurrence of complex noun phrases (these being a recognized feature of academic text) G5: the mean number of words before the main verb in sentences (structurally opaque texts tending to have proportionately more higher order syntactic constituents and great number of words before the main verb)

	G6: logical operators incidence score: include and or, negotiations and a number of conditionals > density means difficult > hypothesizing and linking ideas> a predictor of text adaptation p.198 ***Note:*** In Coh-Metrix paragraph length (number of sentences per paragraph index 4 G2) is measured using an algorithm based on natural language processor from open source library (Grok) where a paragraph is delimited by the number of hard return symbols counted by a text word processor. Index 6 G1: The average number of words per sentence is based on part of speech counted by the Charniak parser. (from Coh-Metrix guidebook)
6 Lexical resources	√ **adopted from Green et al. (2010)** • Word length • Lexical density • Frequency levels V1: *average number of characters per word*), this being a crude indicator of reading difficulty V3 *lexical density* (number of content words as a proportion of the number of grammatical words) and word frequency levels V4, V5 and V6 being the percentage of words occurring among the most frequent and the second and third most frequent

	1000 words in the British National Corpus (BNC) V7 represents the percentage falling outside the 15,000 word frequency level (likely to be technical words or proper nouns) V8 the percentage of words in a text also appearing on the AWL (sub-technical vocabulary) measured V2 *Standardized type-token ratio* (TTR – the ratio of types or different words to tokens: the total number of words occurring in the text) V9: Average number of higher-level constituents per word in the text >>based on semantic hierarchy>>words with more higher-level constituents are more specific – academic texts have more specific terms **Vocabulary** Wordsmith used to calculate TTRs based on 250-words sections of text. (TTR- Standardized type – token ration) the ratio of types or different words to tokens: the total number of words occurring in the text)= percentage / The higher the TTR, the more demanding the passage is likely to be as TTR affected by text length, it is generally recommended

	that standard length to be sued (250)
7 Nature of information √ Text abstractness (Weir, 2005, p.138) Is the text concrete or abstract? Range Whether concrete or abstract (1-5) To measure the abstractness of verbs.	*Note*: (A4) their hypernym values are taken into account. The hypernymity of a verb is measured in terms of a conceptual hierarchy in which the number of levels of superordinate levels above and the number of subordinate levels below the verb being measured are taken into account. This is an indicator of the level of concreteness of abstractness of the verb is measured (McNamara et al., 2012; McNamara et al., 2005).
8 Content knowledge √ Is the topic of the text of general interest or does it require subject specific knowledge on the part of the reader? Range from general to specific (1-5) Cultural background: Is the topic of the text culture-neutral or is it loaded with specific cultural content? Range from cultural neutral to cultural specific (1-5). Language background: The text is significantly easier to understand for readers from a specific first language background. Range from strongly agree to strongly disagree (1-5). Religion knowledge: Is the topic of the text religion-neutral or	

is it loaded with specific religious content? Range from religion neutral to religion specific (1-5). (Weir et al., 2009, p.138)	
9 Response method √ The test response method format is likely to affect the test performance? Range from strongly agree to strongly disagree (1-5). The test tasks provide a variety of response methods Range from strongly agree to strongly disagree (1-5) (Weir, 2005, p.63)	
10 Weighting √ In general, the weighting for different test components are… Range from justified to not justified (1-5). (Weir, 2005, p.64)	
11 Knowledge of criteria √ The criteria to be used in the marking of the test for the candidates and the markers are…Range from explicit to not explicit (1-5). (Weir, 2005, p.63)	
12 Order of items The items and tasks in the test are presented in a justifiable order. Range from justifiable to not justifiable (1-5). √ (Weir, 2005, p.65)	

13 Channel of presentation √ The channel for the target situation requirements of the students being tested is…Range from appropriate to not appropriate (1-5). (Weir, 2005, p.73)	
14 Text length √ The text length for the target situation requirements of the students being tested is…Range from appropriate to not appropriate (1-5). (Weir, 2005, p.74)	
15 Time constraints √ The test time of 40 minutes for the test (e.g. preparation and completion) is...Range from appropriate to not appropriate (1-5). (Weir, 2005, p.68)	

Variables adopted from (Weir, 2005; Weir et al., 2009; Green et al., 2010)

To facilitate judges in making their evaluation, a checklist was drawn up which they could use in making their evaluations (see Appendix 4 for first draft checklist).

The guidelines for constructing this checklist and framing the questions were similar to the ones used for the students' questionnaire in Section (3.4.1) above which followed (Oppenheim, 1992; Gillham, 2000; Brown, 2001; Nardi, 2003; Procter, 2008; Simmonse, 2008 and Wisker, 2008) and in particular (Bachman , 2004; Weir, 2005; Weir et al., 2009). To assist the judges in their task, examples and definitions of the particular features being assessed were included following the advice in Bachman (2004).

3.5.2 Sampling

In this section, sampling strategies for text analysis and expert judges are presented and discussed. A selection of 29 texts was chosen from first year texts books used in Ibra College of Technology and in the Higher College of Technology in Muscat in Oman. The academic texts were selected to be representative of first year programs across all faculties (see Appendix 4). Each of the 29 text extracts was evaluated by three different judges resulting in an overall sample size of 87. The extracts were approximately 500 words in length to correspond with text length of the test passages. The selected texts were taken from different departments as presented in Table (3.4):

Table 3.4 FYA Sample texts

Department/specialization	**Number of courses for first year**
Engineering	8
Business Studies	8
Information Technology	12
TOTAL	29 courses with the addition of a shared course (English Technical writing 1 & 2)

Also, a selection of recent reading test tasks was prepared based on the Level Exit Exam (LEE). Test passages were selected from previous test papers from 2012 to the present. In 2012 the format of the test changed substantially rendering it pointless to include tests earlier than this date. This resulted in a sample size of 5 as shown in Table 3.5:

Table 3.5 LEE Sample tests used in this research

	Test passages Academic year	Semester (1, 2, 3)	Text Length (approximate word count)
1	2012-2013	1	544
2	2012-2013	2	503
3	2012-2013	3	584
4	2013-2014	1	558
5	2013-2014	2	518

The tests were internally created by the teaching staff but following general guidelines governing the design of test tasks followed by all the Colleges of Technology. The reading texts contained approximately 500 words each and comprehension was tested by approximately 25 questions. It is these texts that were loaded into the analytical software.

Identifying and selecting expert judges was based on their qualifications and experience in the general field of L2 teaching and learning, linguistics and assessment, from researchers, curriculum developers, language teachers, and language testers. As a minimum, a qualification of Masters or equivalent was set. In addition, a pedagogical qualification (e.g. PGCE or TESOL) was desirable. A minimum of five years of teaching or lecturing in adult education was also required. A total of 30 judges participated in this part of the research.

In the selected Colleges of Technology, the Heads of Departments for each of the three faculties (Engineering, IT, Business Studies) were requested to supply sample passages from the standard textbooks representing each module of the first year programme. The function of the judges was to rate the various contextual features that were not directly amenable to scalar measurement using a *Likert* scale following a prepared checklist as discussed in the subsection above. Texts from test tasks of the foundation program (LEE) were matched with texts taken from first year academic (FYA) textbooks. The mean response rates from the test tasks and the academic texts were then compared and a statistical test for equality of means was carried out.

3.5.3 Limitations

In this subsection, some limitations of the methods are discussed.

The sample size of the test tasks being used was relatively small. However, it has been explained that, since 2012, the format of the test has been completely altered and that it would have been pointless to include test samples predating the 2012 changes. Accordingly, there were only 5 available tests since the changes of 2012.

The sample size of 87 academic texts may also appear as a limitation. It could be argued that the sample could have been larger; however, this would have placed a great burden on the judges who were giving their services and their time free of charge. 87 texts would amount to reading and evaluating texts of 43,500 total word-count. In view of this, the sample size seemed to be appropriate and adequate to the task.

The variables used to measure some of the features may seem questionable, for example, the number of characters per word as a measure of word complexity. However, the various instruments have been empirically validated and found to be reliable.

Another limitation is that the judgements of experts could be inconsistent due to the degree of subjectivity involved (Bachman, 2004). However, it was likely that highly experienced professional judges would collectively reach decisions that could be deemed to be trustworthy and dependable, as each text/passage was independently examined by three different judges. Having a relatively large sample of judges, 30 in this study, increased the likelihood that an extreme view taken by one judge would be balanced by the ratings of other judges. This process is referred to by Derrida (2004) as “differance” which, in French, is a play on the word which can mean both opposition as well as deferring: “difference as the process of differing-deferral requires us to renounce a logocentric (or epistemic) conception of truth and admit the possibility of that which might always surpass the limits of our knowledge at any given time…there is no extra-linguistical reality by which our various

statements, hypotheses, predictions, etc. might ultimately be assessed in point of their truth or falsehood" (Norris, cited in Derrida, 2004, p. xxxv). Thus, *differance* explains how consensus can be reached among people of differing viewpoints. Inter-subjectivity overcomes the effect of a single extreme subjective view. So the larger number of judges in this study increases the reliability of the findings. Furthermore, the judges' checklist contained clear instructions and brief definitions of the features being assessed and this should have eliminated any possible misconception thereby minimising the effects of subjectivity (Bachman, 2004). Additionally, training was provided for the judges, which consisted of a briefing of the general aims and objectives of the study and the methods being used to address the research questions. Such a briefing provided an introduction for the judges to their particular tasks and gave them an overview of the roles that they would play in the overall study. It is, therefore, concluded that the judges were appropriately orientated towards their tasks of evaluation with this overview in mind.

3.6 Pilot Studies

The instruments were piloted with a view to obtaining feedback on the suitability of the instrument for measuring what it purported to measure (e.g. Thabane, et al., 2010). In piloting the instruments, care was taken to collect evaluative feedback firstly on the instrument viewed as a whole and then on each section of the instrument. The goal of the piloting was to establish whether the instrument was fit for purpose and to eliminate any component which did not appear to be functioning as intended. Any issues related to misconceptions or misinterpretations based on the instrument itself were reported in the pilot study findings and the necessary amendments were made.

The importance of piloting is outlined by many authors such as Bell (2010) who provided a simple procedure for evaluating instruments in order to remove bugs. This procedure should not only look at how well the instrument statements were structured, but should also consider the administration of

the instrument to check time, efforts, expenses and resources. Further, Bell (2010) sees piloting as important for 'analysis' at initial stages. A checklist for evaluating the instrument based on Bell (2010) was created (see Appendix 5). Piloting was planned following a three stage process:

Stage one: proofreading and editing of the statements ensuring consistency

Stage two: Eliciting feedback from experts using the checklist mentioned above

Stage three: Piloting and pretesting with four expert judges for the text analysis and with L2 students, similar to the target group, for the second method under similar conditions (Green, 1998; Brown, 2001). The main purpose of this stage was to identify how effective the instrument was perceived to be by the respondents in describing the thought processes which accompanied their attempts at answering the test questions. The following points were kept in mind in conducting this stage:

- Spotting ambiguities and problems
- Identifying vague directions and any comments by respondents
- Noticing a recurring pattern in responses which might reveal that the issues were not being addressed by the respondents
- Identifying unanswered statements (particularly the ones that many respondents left blank)
- Identifying redundancies
- Deciding whether the instrument served its design purpose (Brown, 2001, pp.62-63).

3.6.1 Pilot study for Verbal Protocol Analysis

Three people with expertise in the teaching of English as a second language were asked to read and evaluate the questionnaire for the VPA. Following the feedback received from them about the questionnaire, certain amendments were suggested (see Appendix 6 for student questionnaire draft II). Questions asking students to nominate their IELTS band had been included in the original design but, following piloting, it was decided that

responses to these questions would really not provide information which would be useful for the research. Most of the students had not taken an IELTS test so that the questions would have been superfluous.

One point which emerged from piloting was that the number of statements was considered to be too large for students to answer directly after taking part in a simulated exam. From the original, the following statements were deleted especially where it was found that statements were repetitive.

- Question 8 should have been awarded more than 1 mark.
- From my geographical knowledge I already understood the role of 'coral reef' in question 10.
- My previous knowledge of the meaning of most of the words in the box in (PART F) meant that I was able to answer this question without referring to the passage.
- In (PART F) it was not necessary to understand the exact meaning of words such as 'culture', 'environment', 'tropical' and 'humid' in order to complete the paragraph.
- The use of 'At the moment', 'but' in lines 33 and 34 helped me to understand the present tourist situation in Zanzibar and future projections.
- It was necessary to refer to the passage to work out the meaning of "prosperous" in line 17.
- The knowledge of the number of marks given to each item affected my time planning and execution
- Before starting to answer the questions it was important to plan how much time should be spent on each item.
- The length of the passage made it difficult to read and understand in the time provided.
- This article is probably taken from a secondary school history book.
- The removal of all these items resulted in a set of statements of more manageable proportions.

It was also suggested that, in their present format, the questions were rather impersonal so it was decided to rephrase all the remaining questions so that they appeared to be more personalised. For example, 'Most of the words on

the passages were concrete, meaning factual straight words and easy to understand, e.g. car, apple rather than abstract words e.g. economy' was replaced by 'I found most of the words on the passages were factual e.g. car, apple'. Certain technical terms were found to be unsuitable such as 'comprehend', 'concrete' and 'abstract'. This resulted in a much more simplified set of statements.

Trialling of this modified VPA instrument was conducted on (Wednesday 26th March 2014) with 12 second language students. The participating students belonged to a group in Bradford University who were improving their English for academic purposes. Although the group was not exactly identical with the students in the Omani colleges, they nevertheless needed to study additional English in order that their standard of English would improve so that they could read, write and speak as appropriate for academic purposes. Table 3.6 presents the background of the twelve students involved in the trial[3]:

[3] The columns in this table were included to show the age, nationalities, language and IELTS attained for these students however the categories are not necessary in the final table for the Omani context as much of this information is irrelevant to the research in Oman (e.g. all the students are Omani).

Table 3.6 Background information of the participants in the trial

Participants	1 Gender	2 Age	3 Nationality	4 School	5 Course	6 Level	7 First language	8 IELTS Reading Band
01	1 male	42	Iraqi	1 Engineering and Informatics	Mechanical Engineering	4 Doctorate	Arabic	5.5
02	2 Female	22	Chinese	5 Social and International Studies	Economics	2 Bachelor	Chinese	5
03	1 Male	26	Iraqi	1 Engineering and Informatics	Computer Sciences	4 Doctorate	Arabic	4.5
04	1 Male	33	Libyan	1 Engineering and informatics	Computing	4 Doctorate	Arabic	5
05	1 Male	24	Chinese	1 Engineering and Informatics	Art Media	3 Master	Chinese	6
06	2 Male	38	Libyan	1 Engineering and Informatics	Communication Engineering	4 Doctorate	Arabic	5
07	2 Male	20	Japanese	5 Social and International Studies	Peace Studies	2 Bachelor	Japanese	5.5
08	2 Female	33	Kuwaiti	Health Studies	Advanced Practice Nursing	Master	Arabic	5
09	1 Male	26	Taiwanese	4 Management	Management	Master	Chinese	5.5
10	1 Male	34	Libyan	1 Engineering and Informatics	Mechanical Engineering	3 Master	Arabic	5.5
11	1 Male	25	Chinese	1 Engineering and Informatics	Digital Arts and Media	3 Master	Chinese	5
12	1 Male	21	Chinese	Engineering and Informatics	Electrical Engineering	2 Bachelor	Chinese	5

All of the students completed the simulated exam within 30 minutes which was within the allotted time of 40 minutes. When it came to the questionnaire, they managed to complete in about 15 minutes. This was, in itself, a good indicator that the statements were clear and unambiguous. It also indicated that they had given their immediate responses to the statements as was desired, rather than spending too much time filtering their responses. Thus, it was considered likely that their more instant responses would also have been likely to be authentic rather than trying to work out responses that they considered to be expected ones. This bore out the importance of the rubric in the questionnaire which stated that there were no correct or incorrect responses to these statements. Thus, there was confidence that the responses to the statements were genuine and authentic.

Each of the participants in the trialling responded to the statements by selecting from a four point *Likert* scale. The absence of a neutral option did not seem to present any problems. Had there been any difficulty for them in not having a neutral option, then they would have skipped some of the questions without giving a reply as none of the four options would have represented their true positions. This did not happen even in one single case, hence it can be safely concluded that the range of answers without a neutral option was appropriate. They answered well within the time scale which suggests that the four available options covered the entire domain of their responses and that time was not spent hesitating. As an additional check, the participants were asked at the conclusion of the trialling whether they had encountered any difficulties with any of the questions and they all concurred that it had been straightforward and easy to answer. Thus, it was safely assumed from the trialling that the instrument was fit for purpose and found to work in practice.

Table 3.7 shows each individual's response to the 34 statements along with a column showing the mean response and the standard deviation:

Table 3.7 Participants' responses to the trialled questionnaire

Statements	Participants													
	1	2	3	4	5	6	7	8	9	10	11	12	Mean	S.D.
S 1	4	3	4	3	4	3	3	3	2	4	2	3	3.166666667	0.717740563
S 2	3	2	3	3	2	2	2	3	3	2	1	3	2.416666667	0.668557923
S 3	3	3	3	2	3	2	2	3	2	3	1	2	2.416666667	0.668557923
S 4	1	4	4	4	3	4	4	3	4	3	4	3	3.416666667	0.900336637
S 5	4	4	4	1	3	3	3	4	4	4	4	3	3.416666667	0.900336637
S 6	4	1	2	1	2	2	2	2	1	2	1	1	1.75	0.866025404
S 7	4	2	3	2	2	1	3	3	3	3	3	3	2.666666667	0.778498944
S 8	3	2	4	2	2	2	4	3	2	3	4	3	2.833333333	0.8348471 1
S 9	2	2	3	2	2	3	3	3	3	3	2	3	2.583333333	0.514928651
S 10	4	2	4	2	2	3	2	3	3	3	2	2	2.666666667	0.778498944
S 11	2	2	4	2	2	3	4	3	3	3	3	2	2.75	0.753778361
S 12	1	2	4	2	2	2	2	4	3	4	2	3	2.583333333	0.9962049 2
S 13	4	4	4	2	2	3	4	3	3	3	4	3	3.25	0.753778361
S 14	4	2	1	3	3	3	4	4	3	3	1	2	2.75	1.055289706
S 15	4	4	4	4	2	2	4	4	4	2	3	3	3.333333333	0.887625365
S 16	2	1	3	2	1	2	2	2	2	3	3	3	2.166666667	0.717740563
S 17	4	4	3	3	4	1	4	4	2	3	4	3	3.25	0.965307299
S 18	4	3	3	2	4	1	4	3	2	3	2	2	2.75	0.965307299
S 19	4	3	4	4	3	3	3	3	3	3	3	3	3.25	0.452267017

S 20	3	3	3	3	3	3	3	4	2	3	3	3	3	0.426401433
S 21	4	3	4	4	3	4	4	4	3	4	3	4	3.666666667	0.492365964
S 22	3	4	4	3	3	4	4	4	3	4	3	3	3.5	0.522232968
S 23	4	3	4	2	2	3	3	3	3	4	2	2	2.916666667	0.792961461
S 24	4	3	3	4	3	3	3	3	3	3	4	3	3.25	0.452267017
S 25	4	3	4	4	4	3	4	3	2	4	3	3	3.416666667	0.668557923
S 26	3	3	4	2	3	2	4	4	3	2	4	4	3.166666667	0.83484711
S 27	4	1	4	3	2	3	4	4	2	4	4	3	3.166666667	1.029857301
S 28	4	4	3	3	3	3	4	4	4	4	3	3	3.5	0.522232968
S 29	4	4	3	4	3	4	3	4	3	4	3	3	3.5	0.522232968
S 30	1	2	2	2	2	3	4	3	3	1	3	2	2.333333333	0.887625365
S 31	4	3	3	3	3	3	4	4	3	4	2	3	3.25	0.621581561
S 32	4	3	2	3	1	3	3	3	3	3	3	3	2.833333333	0.717740563
S 33	1	1	3	3	2	4	3	2	3	2	3	3	2.5	0.904534034
S 34	4	3	3	3	3	2	4	3	2	3	3	3	3	0.603022689

4= strongly agree, 3= agree, 2= disagree, 1= strongly disagree

In general, there did not appear to be a pattern in the responses that would have indicated the possibility that particular students had always opted for the same response raising suspicion that they might not have been responding in an honest fashion. The column with mean and standard deviation is included but not with the intention of analysing the responses as this was simply a trial to look for inconsistencies or anything which might have shown further ambiguities in the statements. About one third of the responses showed a fairly high standard deviation close to or above 1 but

this was not a concern as such a variation from the mean might well have reflected the range of responses to the various statements . Statement 20 *(I think the number of marks given for each question was appropriate)* received ten responses of 3, the other two being 2 and 4 which gave a mean response rate of exactly 3 (equal to *agree*). However, this clustering of similar responses to this question is not considered to be problematic as it was the only instance of such clustering in all of the 34 statements and was therefore considered to be a true reflection of the students' responses to the statements which were phrased quite unambiguously.

Of more concern would have been clustering by participants, meaning that a particular participant tended to give the same response to each statement. However, a visual check showed that no such clustering existed and that the trial had indicated that the instrument was fit for purpose and ready to be used in the actual research.

Reflection on the trialling, however, focused on the fact that these participants, although L2 students, were studying in a context where English was the spoken language. Their average IELTS score was 5.2 with only one participant as low as 4.5. Additionally, the majority of the participants were at Masters or PhD level whereas the students in Oman were more likely to have an IELTS score of around 4. It was therefore considered that, in order to facilitate the Omani students, the statements would be presented both in English and Arabic. The final version of the questionnaire is presented in (see Appendix 7).

3.6.2 Pilot study for Automated Text Analysis and expert judges

At first it was proposed to have one single checklist to cover both types of texts (test tasks and academic extracts). However, this might have been confusing in that it would require rubrics instructing judges not to answer certain questions. Therefore, it was considered preferable to have two separate checklists for evaluating each type of texts. Certain items such as (Task demands) were not applicable to academic texts. So, for that reason, it was considered to be less complicated to create two separate checklists (see

Appendix 8). Using two checklists in the trialling seemed to work very well and for that reason it was adopted for the final version.

Another consideration was the size of the checklist. However desirable it might have been to include other matters in the checklist, the overall aim was to keep the checklist as brief as possible as the judges would have significant amounts of readings to do and they were giving their time and expertise free of charge. Nevertheless, the checklist was comprehensive and covered all of the variables that were desired to be tested in this study.

Generally, keeping the checklist as brief as possible meant that there would be one statement for each variable. However, some of the features being tested did require more than one statement as they contained a number of subsets (e.g. for discourse mode, there were four statements embracing each of four features: genre, rhetorical task, pattern of exposition and rhetorical organisation; two statements for grammatical resources: grammar and cohesion; and three statements for content knowledge: general, cultural, language and religion).

Statements (13, 15 & 16) were adjusted in the light of feedback during piloting so that a five point scale could be applied in a way that made more sense in view of how the statements had been previously framed.

From: (13) The text is significantly easier to understand for readers from a specific first language background

To: Is the text first language background neutral or first language background specific?

Options: (1 first language background neutral 2 3 4 5 first language background specific)

From: (15) The test response method format is likely to affect the test performance?

To: How likely is the test response method to affect the test performance?

Options: (1 likely 2 3 4 5 unlikely)

From: (16) The test tasks provide a variety of response methods

To: The test tasks provides…

Options: (1 a variety of response methods 2 3 4 5 no variety of response methods)

More importantly, as these statements were originally phrased, there was a danger that they might skew it more in favour of the judges' own feelings or opinions rather than being a more objective statement based on the text itself. However, this is not to imply that the judges would misunderstand what was being asked of them as the very reason for inviting them in the first place was the fact that they were experts and would be expected to give a fair opinion on the matter. Simply put, the reason for reframing statements (13, 15 and 16) was that they would be aligned more closely to the original research questions.

In the trialling stage, four expert judges were provided with a sample of two texts: one test task and one extract from academic level. The two sample texts that were trialled are presented in Table 3.8:

Table 3.8 Sample trialled texts

Text type version	Department	Course	Year	Title	Text length
Test Task	English language Centre	Level Four Exit Exam	Foundation year	Zanzibar	503 words
Extract Text	Business Studies	Introduction to Business	First year academic	Corporations	494 words

The participating judges and their academic backgrounds are presented in Table 3.9:

Table 3.9 Participants' background information

Expert Judge	Qualification	Range of experience	Years of experience in teaching English language	Location
PS	/	Lecturer of English as a second language	Not provided but it is assumed as having a minimum qualifications as usually required by the University of Bradford	University of Bradford, English Language Centre, UK
EN	BS in Education (English)	Lecturer of English as a second language	Over 10 years of teaching English to adults at tertiary level	Oman, Ibra College of Technology, Language Centre
DF	MA in English Language Studies & Linguistics	Lecturer of English as a second language	Over 10 years of teaching English to adults at tertiary level	Oman, Ibra College of Technology, Language Centre
CO	M.B.A., International Business Development, Manager	Taught English in different Gulf states including Oman	Over 5 years of teaching English as L2 to adults at tertiary level and administering	Qatar

The participants were selected and identified according to the conditions set out in the design and content subsection above. The four participants had many years of experience in teaching and assessment of English for second

language students at tertiary level similar to the target situation of this thesis. Two of them were non-native English speakers (EN & DF) and the two others were native (PS & CO).

The checklists and texts were sent by email; no training had been provided although this was provided in the actual research. The reason that no training was provided was due to the distance involved and the difficulty of providing training by email. In the research itself, there was a meeting with the judges where training was provided and the rationale of the study was explained with opportunities for the judges to raise any questions. Nevertheless, in the trialling, the participants expressed no difficulty in answering the questions even without training and it was obvious that the instructions had been clearly stated.

Of the four participants in the trialling, only two returned their responses which were, nevertheless, considered to be sufficient for the purpose of trialling, as two people with expertise were giving their assessment of the instrument.

Table 3.10 summarises responses from both experts in evaluating the instrument:

Table 3.10 Expert judges' instrument piloting: Checklist evaluation sheet

1. How long did it take you to complete?
EN: 30 minutes **DF:** Maybe 10-15 minutes each check list
2. Were the instructions clear?
EN: Yes **DF:** yes
3. Were any of the questions unclear or ambiguous? If so, will you say which and why?
EN: No **DF:** I understood all the questions except on the degrees of choices. Please see my comments earlier.
4. Did you object to answering any of the questions?
EN: No **DF:** No
5. In your opinion, has any major issue been omitted? Please specify.
EN: Yes. Levels of questioning, order of tasks, and face validity of the test should be included in the checklist. **DF:** None reported
6. Was the layout of the questionnaire clear/attractive?
EN: Yes **DF:** I have sent the comments for the Test of the Zanzibar reading test.
7. Any comments?

EN: Nothing.

DF: Please refer to the comments I've sent earlier.

For the Expert Judges Draft 1 - Test Task Evaluation:

- The "Text" is appropriate.
- Rhetorical Task: (page 3, #4)
  - Suggestion: Options may be **descriptive, narrative, expository, argumentative**. (*This may also be applicable to the other Checklist.)*
- Response Methods: (page 5)
  - There is a variety of response methods. However, were they appropriately developed? *This is one point to be considered in finalizing the Judgment Checklist.*
- Weighting: (page 5, #17)
  - Answer is based on the assumption that test construction was based on a Table of Specifications (TOS).
- In reference to the General Foundation Programme expected learning outcomes, the test failed to measure the achievement of all the expected learning outcomes.
  - Suggestion: Levels of questioning be included in the *Judgment Checklist*.

With regard to the time taken to compete both tasks (Q1 above), both experts seem to take similar times. On the question of the clarity of the instructions, both replied that these were clear (Q2). On the question of ambiguities (Q3), one expert questioned the range of choices; however, this issue was later addressed during the training of the experts in the actual research. Regarding having any objections to any of the questions, both replied that they did not have any objections (Q4). The fifth question asked if they thought that any major issue had been omitted; one of the respondents stated that levels of questioning, order of tasks, and face validity of the test had not been included in the checklist. This was correct. However, these three issues lay outside of the scope of the current research which was based on the representativeness of the content of the text. But as has already been stated, during training, the full rationale of the research was explained to the expert judges in the actual research. From the responses it would appear that the layout of the questionnaire made it easy to use. Under *any comments*, a few issues were raised by one of the experts. One was concerned with various types of texts (e.g. descriptive, narrative…etc.). However, these had actually been included but, somehow, this respondent seemed to have missed them. A second comment concerned the variety of response methods and questioned whether they had been appropriately

developed. This was also a valid question in general but, again, it raised an issue that lay beyond the scope of the research as mentioned previously. Training of experts actually helped to clarify the focus of the current research and issues raised in the piloting were borne in mind during the actual training given in the research proper. The final comment raised by this respondent concerned the learning outcomes being measured by the test. Again, while this was a valid question in its own right, it was properly a matter for another research and lay beyond the scope of the current research. Nevertheless, these three comments alerted the researcher to the need for providing expert judges with some training and a clear rationale for the current research so that they would understand what issues lay within the remit of the study and what issues were beyond its scope. Thus, the trialling with the two expert judges provided the researcher with some insights into what to include in the training for the expert judges in the actual research.

Nevertheless, the general impression from the expert judges' feedback was that the checklists were clear, unambiguous and were relatively easy to complete in a short time. The expert judges' responses did confirm one issue in the other trial, which related to the phrasing of the statement so that it could be appropriately reflected on the *Likert* scale. This issue has been addressed especially in relation to questions (13, 15 and 16) and adjustments have been made to the final version, which is in Appendix 8.

3.7 Validity and reliability

Despite the concern raised by Ericsson (2006) regarding retrospective techniques such as have been employed here, it was, nevertheless, considered that validity would be assured by means of the following:

1. Avoidance of qualitative type and more open-ended statements where generalised multiple-variable responses might be received. Under investigation here was the cognitive processing involved in taking a reading test. Qualitative statements of a more open type: "Although this can contribute to rich theory building, it may not address aspects of cognitive processing that are of primary interest to the investigator"

(Pressley and Afflerbach, 1995, p.11). Accordingly, great attention was paid in formulating the statements to ensure that they conformed to the following criteria:

- Focusing on single variables (recommended by Ericsson, 2006).
- Allowing inter- and intra-reliability comparisons for reliability (Seliger and Shohamy, 1989).
- Minimising participants' writing.
- Presenting brief and clear statements with *Likert* scale (4 instead of 5 to minimise ambiguity and thus achieve higher objectivity).

2. Recall took place without delay immediately after the reading task (see Cohen, 2000, p.138).
3. Clear instructions and guidance were given with the possibility of questions being asked by the respondents if further clarification was required.
4. The groups were relatively small consisting of about 30 participants and this meant that individual attention could be given.
5. A robust piloting process eliminated ambiguities in how the statements were framed and also the removal of language which might have been too technical. Moreover, the insights of experts in the piloting improved validity.
6. Voluntary participation was on a random basis. This was believed to "enable the findings to have greater generalisability (external validity), i.e. to represent the wider population (Cohen et al., 2011, p.326).

3.8 Ethical issues

The main ethical issues involved were concerned with the sample of students and assurances of good practice. The following were the most important considerations which the researcher has taken into account in order to ensure high standards and procedures of ethics were applied as recommended in the literature (e.g. Sture, 2010):

- **Identifying, approaching and recruiting potential participants in the project**

Since he is in direct charge of all Colleges of Technology, the Director General of Technological Education (DG) was the first contact. He was contacted by telephone and permission to proceed with the research was confirmed by e-mail. Having obtained the approval, emails were sent to deans and heads of departments, informing them of the purpose of the research and how much its success would depend on their wholehearted cooperation (see Preliminary letter in Appendix 9). After initial approval, they received a form of consent (see Appendix 10) and the instrument. However, respondents were informed of their right to have their data withdrawn from the raw data within two weeks should they so desire.

Participation was on a voluntary basis but it was anticipated that heads of departments would promote the project and present it in a good light in view of the potential benefits which were eventually expected. Participants were given a simple explanation of the research and guidance.

- **Any potential for physical and/or psychological harm / distress to participants arising from their taking part in your research and planning to minimise it**

Piloting revealed that there were no such issues likely to cause physical or psychological harm to the participants. The nature of the phenomena under investigation was not such as would entail emotional involvement or distress on the part of participants. Students were assured that the test was an experiment and that the outcome of the test would have no bearing on their future development in the college. It did however give the students an additional opportunity to practise their reading skills and to reflect on the thought processes involved in completing the various task as discussed earlier.

- **Measures put in place to ensure confidentiality and/or anonymity of personal data**

The measures included the following:

- ❖ Explicit written or verbal assurance explaining confidentiality in the requested consent and questionnaire.
- ❖ Names of participants or their personal details were not included and details of gender and subject specialisms were kept on a separate detachable sheet. This ensured confidentiality and the elimination of any possibility of individuals being identified.
- ❖ Collected data was not shared with other researchers except in summary form and with maximum care. Students were assured that participation in an honest fashion would have no detrimental consequences for them as the responses would be presented in a generalised way.
- ❖ Disposal of records would be achieved by means compliant with the University's requirements.
- ❖ For greater confidentiality, where there was any risk that participants might be easily identified, findings were generalized rather than individually presented.
- ❖ Data were kept in a locked secured place and data stored electronically were password protected.

Although no risks were anticipated, such measures were necessary to protect data from loss or harmful use. Additionally, these measures were a protection for the supervisors, researcher, sponsors and the university from any unexpected or unforeseen risks.

Finally, as a matter of courtesy, all participants received a message of gratitude at the end of the research and a note of acknowledgement was included in the foreword to the research.

3.9 Practical considerations

Finally, practical challenges of the design were considered:

- Time: having multiple phases would add to the time involved in data collection and analysis (Creswell, 2009; Robson, 2011).
- Analysis: the issue of a mismatch of findings was not considered to be of great concern for this study as data from one method was not being collected in a mixed method style in order to corroborate the findings of

the other method. Rather, it was expected that the findings of the VPA method would serve to amplify the findings from the text analysis and expert judges' phase. A potential issue was considered whereby the expert judges' assessment of advanced students in English might not correspond closely with the situation in Oman.

- Logistical: collecting the verbal protocol data necessitated travelling to Oman and this involved considerable planning to take into account timings of semesters and holidays. Additionally, the researcher's family life and the possibility of unforeseen circumstances had to be considered.
- Language: It was considered necessary to present the questionnaire to the students in Arabic as well as in English.

3.10 Summary

This chapter has discussed issues of research design and methodology. Although the research was largely quantitative in nature, a qualitative approach based on expert judgements was used to collect data on contextual features which were not directly amenable to quantitative measurement; however quantitative analytical methods were applied throughout (see Chapter 6 and Chapter 7). Piloting and trialling were most important for this study and led to the development of instruments which were considered to be valid and fit for purpose. One aspect of the analysis of the data was concerned with the robustness of the variables gleaned from Khalifa and Weir's (2009) model (as discussed in Chapter 2 Section 2.5). It was proposed that correlation tests would be conducted to establish that each variable was measuring distinct features and to discover whether there was any significant overlap between the variables. Finally, a factor analysis was also considered appropriate whereby, taking into account the students' actual test scores, it would be possible to identify the actual processes used against the intended processes to be tested for each variable. In this way, it was expected that factor analysis would provide cogent evidence for answering the research questions and to establish the cognitive processes the students actually used. These considerations are discussed in depth in the following chapters (see Chapter 4 and Chapter 5).

Chapter 4 Data collection and analysis I: A natural experiment utilising Verbal Protocol Analysis (VPA)

4.1 Introduction

This chapter presents the results and findings of the data which were collected to answer the first research question:

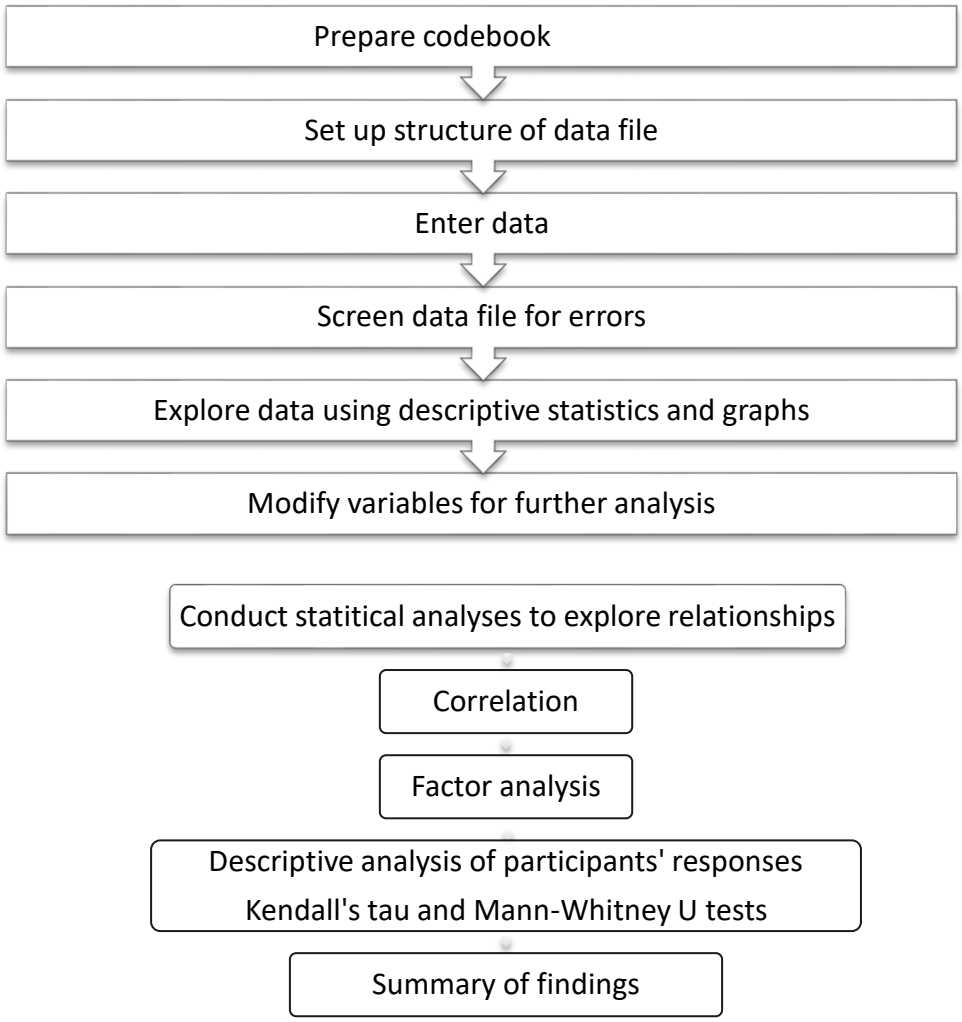
"What are the cognitive processes by which students engage with the texts and tasks in reading tests?"

In a natural experiment in the form of a simulated test (see Chapter 3 Section 3.4), data was collected from two colleges, Ibra College of Technology (ICT), which is a regionally based college, and the Higher College of Technology (HCT), located in Muscat, the capital city of Oman, which represents a more urbanized metropolitan context. There was a total of 202 participants. The conventions of the American Psychological Association (APA) were followed in reporting and presenting the findings (American Psychological Association, 2010). Descriptive statistics were obtained by using IBM Statistics SPSS version 21 to give an overview of the data and inferential statistics were obtained to investigate any significant differences in the data (Frank and Althoen, 1994; Antonius, 2003; Sarantakos, 2007; Procter, 2008). The 34 variables measuring the cognitive processes (see Chapter 2 Section 2.7 and Chapter 3 Section 3.4) were subjected to correlation tests to determine whether each variable measured distinct constructs or features. Additionally, a factor analysis was run to further explore the variables and any associations. Both of these tests served to validate the model of Khalifa and Weir (2009) in two important ways. Firstly, it was important to establish that the 34 variables (see Codebook in Table 4.1) were measuring distinct cognitive features in order to answer the research question which concerned the different cognitive processes used by students in reading and comprehension tests. Secondly, establishing that the variables were measuring distinct cognitive processes was important for avoiding construct underrepresentation and construct irrelevant variance. Of equal importance was the fact that the test content should be a sufficient basis upon which

inferences could be made about the participants' performances. This procedure of reading model validation is supported by a number of authors (see Chapter 2 Section 2.5).

The statements responses from the questionnaire were examined and significant findings ($p < .05$) were noted. This involved the analysis of the results measuring the extent to which the students reported that they were using each of these cognitive processes. Finally, Kendall's tau and, where appropriate, Mann Whitney U tests were run to identify significant differences based on test item scores between those who used the intended process and those who used an alternative process. The entire analytical process is outlined in the flowchart (see Figure 4.1):

Figure 4.1 Data analysis process



Adapted from Pallant (2010, p.28)

4.2 Descriptive statistics

The data from the questionnaire responses were entered into IBM Statistics SPSS version 21 and scoring was applied following the guidance given by many scholars in the field (e.g. Frank and Althoen, 1994; Field, 2001; Wagner, 2010). The participants indicated the strength of their agreement or disagreement with the 34 statements using a 4-point *Likert* scale:

- Strongly Agree = 4,
- Agree = 3,
- Disagree = 2 and
- Strongly Disagree = 1

Variables were entered and the cognitive processes involved in each of the 34 statements were appropriately coded for SPSS processing (see Table 4.1):

Table 4.1 Codebook

Variable/Statement	SPSS variable name	Coding instructions
Part I: Background Information		Number assigned to each survey
College	College	1 = Ibra College of Technology (ICT) 2 = Higher College of Technology (HCT)
Gender	Gender	1 = Male 2 = Female
Subject Specialisation	Subject Specialisation	1 = Engineering 2 = IT 3 = Business Studies 4 = Applied Sciences 5 = Pharmacy 6 = Fashion Design 7 = Photography
Part II: Think Aloud		
1. I was quickly able to find the information required to answer the questions in PART A.	Scanning expeditiously	4 = Strongly Agree 3 = Agree
2. I found it difficult to decide whether to skim (fast read) or read carefully the whole passage in order to answer questions.	Discernment1	2 = Disagree
3. I found it difficult to decide what was the main idea in Paragraph 3 because the sentences were too	Grammatical resources1	1 = Strongly Disagree

	long (question 6).	
4.	I had to read other sentences carefully in addition to the sentence on line 16 in order to answer question 8.	Careful global
5.	I was able to answer question 9 by carefully reading the sentences in line 17 and 18.	Careful local
6.	I did not need to refer to the passage to answer question 8 because I already knew the meaning of ‘extensively’.	Lexical resources 1
7.	I was unsure whether long detailed answers or short answers were required in answering Part D.	Rubric 1
8.	I had to quickly search the whole passage to find the information required to answer question 10.	Expeditious search 1
9.	I had to quickly search the whole passage to find the information required to answer question 11.	Expeditious search 2
10.	I had to quickly search the whole passage to find the information required to answer question 12.	Expeditious search 3
11.	I had to quickly search the whole passage to find the information required to answer question 13.	Expeditious search 4
12.	I understood from question 12 that the Sultan moved the entire city of Muscat brick by brick to Zanzibar.	Functional resource Ideational
13.	To answer questions 18, 19, 20, 21, and 22, I had to read and understand the whole passage.	Response method1
14.	I did not need to know the exact meaning of the words such as ‘culture’, ‘environment’, ‘tropical’ and ‘humid’ in order to complete the paragraph.	Lexical resources 2
15.	I needed to read beyond the sentence in line 7 the sentence in order to understand the meaning of ‘the climate is tropical and humid’.	Grammatical resources 3
16.	I think that the history of Zanzibar is described as ‘colourful’ in line 27 due to the amount of sunshine that Zanzibar receives.	Functional resource imaginative

17. I could easily understand 'it' referred to in line 7	Grammatical resources 2
18. I found it easy to decide which of the three options best express the main idea of Paragraph 1 because the sentences were short.	Grammatical resources 3
19. I found the test instructions easy to understand	Rubric 2
20. I think the number of marks given for each question was appropriate.	Weighting
21. I found the variety of question types (e.g. true/false, short answer, gap filling, etc.) helpful in allowing me to show my skills.	Response method 2
22. It was helpful to know how many marks were allocated to each item.	Knowledge of criteria
23. I thought it was important to answer the questions in the order they were presented.	Order of items
24. A time line of main events would have aided and supported my passage comprehension.	Channel of presentation 1
25. I did not need to read every word of the passage in order to understand its meaning.	Passage length
26. I looked at the questions first before deciding whether to read the passage carefully or quickly.	Discernment 2
27. 40 minutes was sufficient time to answer all the questions.	Time constraint
28. Recognising the overall purpose of the passage helped my understanding.	Overall passage purpose
29. I think the way the paragraphs are arranged helped me to understand the passage.	Discourse mode 1
30. This passage is suitable for Omani students.	Writer-Reader
31. The passage presents the historical development of Zanzibar in an easy - to-understand way.	Discourse mode 2

32. I found most of the words on the passages were factual e.g. car, apple.	Nature of information
33. The article was mostly familiar to me due to my knowledge of the history of the connections between Zanzibar and Oman.	Content knowledge
34. I was easily able to understand the passage even though there were no headings.	Channel of presentation 2

The sequence of these features in the questionnaire was based on the order in which they appeared in the test paper in the natural experiment/questionnaire (see Chapter 3 Section 3.4). The analysis, therefore, follows that order. The following features from Khalifa and Weir's (2009) model of reading were addressed (see Chapter 2 Section 2.3):

- Scanning expeditiously= statement 1
- Discernment= statements 2 and 26
- Examining Careful reading at global level = statement 4
- Examining Careful reading at local level = statement 5
- Rubric = statements 7and 19
- Examining expeditious reading at local level search = statements 8, 9, 10, 11.

The remaining variables comprised the additional features from task setting and linguistic demands (see Chapter 2 Section and Section 2.5). Descriptive statistics were obtained from SPSS and are presented in Table 4.2:

Table 4.2 Descriptive statistics

College		Frequency	Percent
Valid	IbraC	98	49
	MuscatC	104	51
	Total	202	100.0
Gender			
Valid	Male	139	69
	Female	62	31
	Missing	1	.5
	Total	202	100.0
Subject Specialisation			
Valid	Engineering	112	55
	IT	27	13.5
	Business Studies	32	16
	Applied Sciences	27	13.5
	Pharmacy	1	.5
	Photography	3	1.5
	Total	202	100.0

Table 4.2 shows the total number of respondents (*N* = 202) and the numbers from each college; Muscat (*n* = 104) and Ibra (*n* = 98). 69% (*n* = 139) of the respondents were male, 31% (*n* = 62) were female and .5% (*n* = 1) respondent who did not declare their gender. The percentages based on gender closely reflected the gender balance in Omani Colleges of

Technology due to an enrolment policy whereby 70% male and 30% female applicants were accepted (see Table 1.2 Chapter 1 Section 1.1). 55% (*n*= 112) of the respondents had Engineering as their subject. IT, Business Studies, Applied Sciences constituted 13.5% (*n* = 27), 13.5% (*n* = 27) and 16% (*n* = 32) respectively. Only 1.5% (*n*= 3) respondents were photography students and there was just a single pharmacy student (.5%). Consideration was given to these final two subject specialisations with the possibility of removing them from the data but it was decided to include them and to note any SPSS output flagging these up for any reason.

4.3 Correlation results

The next step in the investigation was to examine the 34 variables to check for any associations. Checking the variables for levels of association is recommended by many authors e.g. (Wagner, 2010). This was necessary to discover whether any pair of these variables was associated to such a degree as might suggest overlap or duplication of the context validity features being investigated. A complete correlation matrix (*Bivariate /Pearson*) was generated by SPSS and examined for any significant associations. A number of correlations were found to be significant ($p < .05$) but the associations were, in all cases, found to be weak i.e. ($-.5 < r < +.5$) (e.g. Greasley, 2008). This yielded some evidence for each of the variables being valid measures of distinct cognitive processes and provided confidence in the scheme of variables identified in Khalifa and Weir's (2009) model including their matrix of modes and levels of reading.

4.3.1 Response method 2 and Discourse mode 1

The highest association was between Response method 2 and Discourse mode 1 (Pearson $r = .447$, $n = 202$, $p < .001$ two-tailed). Returning to statement 21: *I found the variety of question types e.g. true/false, short answer, gap filling, etc. helpful in allowing me to show my skills* and statement 29: *I think the way the paragraphs are arranged helped me to understand the passage*, it is clear that, despite some association, distinct

cognitive processes are involved. The former was concerned with the variety of ways the students were given to respond to questions and the latter was more concerned with how sentences and paragraphs aided comprehension through their cohesiveness. There is, therefore, no concern that there is duplication or a high degree of overlap between the two variables.

Discourse mode 1 was also found to be weakly associated with Rubric 2 (Pearson r = .446, n = 202, p < .001 two-tailed). Other pairs of variables showing some degree of correlations included Expeditious 1 and 2, Expeditious 2 and 4, Expeditious 3 and 4, and Discourse mode 1 and 2. It is not surprising that there would be some degree of correlation on variables testing the same contextual feature, although covering different aspects of that feature, but again, the correlations were weak.

4.3.2 Scanning expeditiously and Grammatical resource 2

Other variables showing significant correlation were Scanning expeditiously and Grammatical resource 2 (Pearson r = .423, n = 202, p < 0.001 two-tailed), and Discernment 1 and Grammatical Resource 1 (Pearson r= .426, n = 202, p < .001 two-tailed).

Considering the association of Scanning expeditiously and Grammatical resource 2, these refer to statements 1: *I was quickly able to find the information required to answer the questions in PART A* and statement 17: *I could easily understand what 'it' referred to in line 7* respectively. Both of these variables involved locating specific information at a local level; in the case of scanning, the test taker was asked to find a specific piece of information in order to answer the question and, in the case of grammatical resources, some local scanning was required in order to locate what was referred to by the pronoun "it". This partly explains why there was some association between the two variables.

4.3.3 Discernment 1 and Grammatical resource 1

The next pair of variables where weak associations were found was Discernment 1 (Statement 2: *I found it difficult to decide whether to skim (fast*

read) or read carefully the whole passage in order to answer questions, and Grammatical resource 1 (Statement 3: *I found it difficult to decide what was the main idea in Paragraph 3 because the sentences were too long (question 6)*).

In response to the statement that it was difficult to understand the meaning of the passage because the sentences were too long, over half the students (57%) either strongly agreed or agreed (*n = 40, n = 75*, respectively). It is, therefore, not surprising that this variable was associated with discernment due to the difficulty faced by students at level four (highest level prior to first year academic) in being able to self-manage and cope with sentences containing subordinate clauses.

Overall, the fact that no strong association was found between any pair of the variables was a good indicator of the robustness of this scheme of context validity. It suggested that the variables were distinct from each other and were, therefore, measuring different aspects of the context validity feature being studied. Even among some of the variables where weak correlations were found, this was not unexpected since the particular activity in the test called for more than a single cognitive feature (e.g. Discourse mode 1 and Discourse mode 2). Some variables measured Functional resource such as ideational and imaginative features but yet showed only a weak association, which further attests to the robustness of these variables. Thus, each one of the variables is considered as measuring a distinct cognitive feature and there was no evidence that there was a serious overlap that might suggest the advisability of dropping a particular variable. It was important to conduct this correlation test to confirm that the variables representing the reading construct were measuring distinct variables and the correlation matrix did at least suggest that the variables were representing underlying constructs. The difficulty of arriving at such a set of variables has been pointed out by Alderson and Kremmel (2013) who supported the view that such a procedure was important especially for “an understanding and diagnosis of learners’ strengths and weaknesses in reading in a second language” (Alderson and Kremmel, 2013, p.1) (see Chapter 2 Section 2.5). This confirms the view of Fulcher (1998) referred to earlier (see Chapter 2

Section 2.5) who sees the necessity for model validation as urgent due to the risk to test validity. Validating a model in this way leads to greater confidence on the part of the various stakeholders regarding inferences that can be made about the cognitive processes through which students engaged with texts and tasks in a reading test. It would also strengthen the validity of conclusions being made in addressing the first research question which was to determine the cognitive processes by which students engaged with a reading test.

Having established that the variables were robust and measured distinct features based on the evidence of the correlation matrix, it is important to point to an important caveat in relation to the correlation test. The correlation coefficient is a measure of a linear association between variables; it is therefore possible to have a very strong relationship between variables where it is not modelled on a straight line and in which case *r* could be small even though in reality a strong association existed (The Open University, 1988). Due to this caveat, it was decided to conduct a factor analysis and the findings of this analysis are presented in the following section.

4.4 Factor analysis

As an additional check that the variables were measuring distinct cognitive features, a factor analysis of the variables was conducted. Although factor analysis is usually aimed at the reduction of variables, it was used in this research to discover whether there was sufficient evidence that some variables could be dropped from the model without affecting the model's overall validity (Pallant, 2010).

The procedure outlined in (Pallant, 2010) was followed by first establishing that the data was suitable for factor analysis. Output for the Kaiser-Meyer-Olkin Measure of Sampling Adequacy (KMO) test is presented in Table 4.3:

Table 4.3 KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.752
Bartlett's Test of Sphericity	df	561
	Sig.	.000

Table 4.3 shows a **KMO** measure = .752. This is higher than .6 which is the lower threshold for the suitability of the data for factor analysis (Pallant, 2010). Secondly, **Bartlett's Test of Sphericity** is significant (df = 561, $p <$ .0005). Thus, the two criteria required for assessing the suitability of the data for factor analysis have been satisfied.

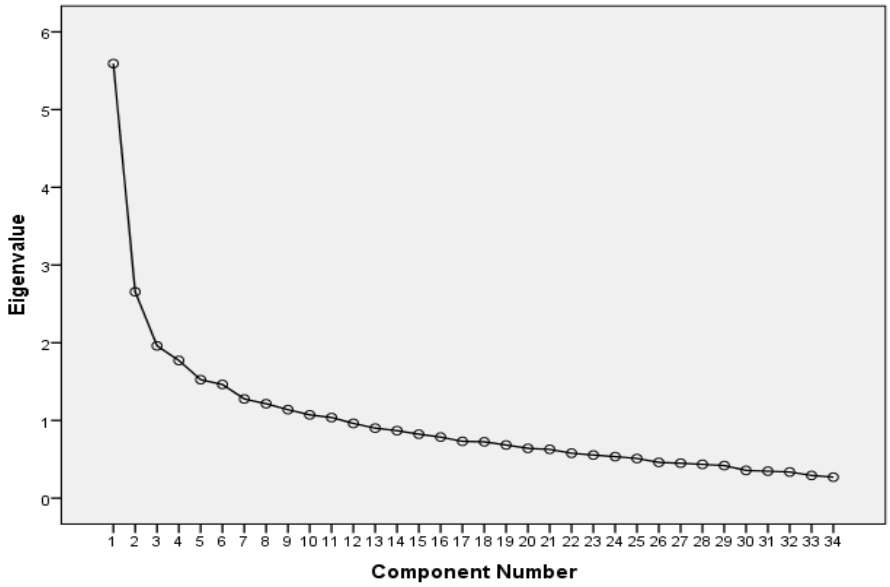
The correlation matrix generated by SPSS was examined and very few correlations coefficients above .3 were found. This is another indicator that factor analysis was justified (Pallant, 2010).

The next step was to determine which factors had Eigenvalues above 1 which would be the criterion for extraction if our aim was to devise a simple explanatory model. This step was necessary to identify those variables which accounted for 60% or more of the variance (Pallant, 2010) (see **Total Variance Explained** Table 4.4 in Appendix 11).

From Table 4.4 only the first 11 variables had Eigenvalues > 1. These 11 variables accounted for 61% of the variance.

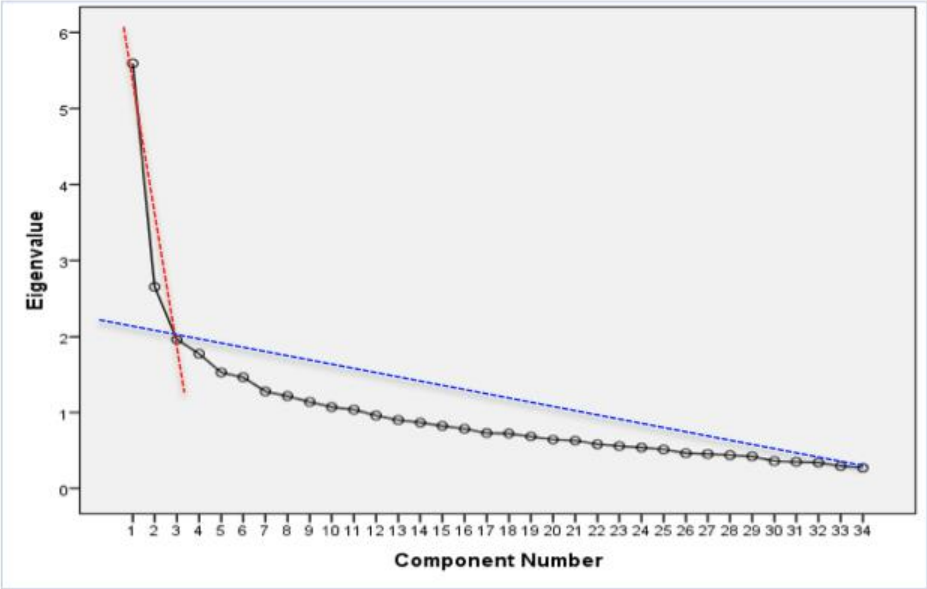
The next step suggested by Pallant (2010) was to obtain and examine a **Scree Plot** (see Figure 4.2):

Figure 4.2 Scree Plot



The Scree Plot was difficult to interpret due to fluctuations beyond the third point, so it was decided to first explore the model based on extracting the first six variables (Pallant, 2010). However, examining the results based on a six component solution was not satisfactory especially as component 2 was clearly expeditious reading but it would be difficult to denote what the other 5 variables were actually measuring. A number of alternative approaches were tried and, in the end, a solution based on the third point as the point of inflexion produced the most satisfactory result which was based on a two component model (see Figure 4.3):

Figure 4.3 Scree Plot



It is only the points before the point of inflexion that determine the appropriate number of factors to be retained (Field, 2012).

Thus, a two factor approach was considered to be the most appropriate solution. The Component Matrix, Pattern Matrix and Component Correlation Matrix were now rerun based on the two components and the outputs from SPSS are presented in (Tables 4.5, 4.6 and 4.7 respectively). This time, the results were more satisfactory for explaining the data and variances. Component 2 can readily be interpreted as Expeditious Reading. However, Component 1 consists of basic reading processes which probably include self-management processes as well as drawing on prior knowledge.

The Component Matrix (Table 4.5) shows that all of the variables loaded on either of the two factors:

Table 4.5 Component Matrix

Statement #		Component	
		1	2
31	Discourse Mode 2	.628	
19	Rubric 2	.625	
29	Discourse Mode 1	.613	
33	Content Knowledge	.607	
21	Response Method 2	.592	
28	Overall Passage Purpose	.571	
32	Nature of Information	.553	
17	Grammatical Resources 2	.542	
30	Writer-Reader	.535	
34	Channel of Presentation 2	.520	
1	Scanning Expeditiously	.519	
20	Weighting	.504	
5	Careful Local	.499	
18	Grammatical Resources 3	.477	
22	Knowledge of Criteria	.457	
25	Passage Length	.437	
4	Careful Global	.407	
27	Time Constraint	.400	
26	Discernment 2	.372	
14	Lexical Resources 2		
23	Order of Items		
12	Functional Resource Ideational		
9	Expeditious Search 2		.630
10	Expeditious Search 3		.622
11	Expeditious Search 4		.595

8	Expeditious Search 1	.495
15	Lexical Resources 3	.482
2	Discernment1	.386
24	Channel of Presentation 1	.348
7	Rubric 1	.315
3	Grammatical Resources1	
13	Response Method1	
6	Lexical Resources 1	
16	Functional Resource Imaginative	

Extraction Method: Principal Component Analysis.

a. 2 components extracted.

However, 7 of the variables have low loadings below the threshold of .3 and so are not showing any values in the table. These are

- Statements 6 and 14: Lexical resources 1 and 2,
- Statement 23: Order of Items,
- Statements 12 and 16: Functional resource ideational and imaginative,
- Statement 3: Grammatical resources 1, and
- Statement 13: Response method 1.

It must be remembered that some features have been represented more than once by test items. Taking this into consideration reduces the list of low loading variables to just two variables which are Order of items and Functional resources. Both of these are discussed in Section 4.5. The other 5 variables that have low loadings were represented under another variable, for example Response method is also represented under Response method 2 loading (.592) on Component 1.

The Correlation Matrix conducted in Section 4.3 above did not provide evidence of duplication and, as a consequence, argued for the retention of

all 34 variables. However, the Pattern Matrix (Table 4.6) shows the same 7 items as low loading on either component as above.

Table 4.6 Pattern Matrix

Statement #		Component	
		1	2
19	Rubric 2	.662	
31	Discourse Mode 2	.653	
29	Discourse Mode 1	.613	
33	Content Knowledge	.593	
21	Response Method 2	.587	
1	Scanning Expeditiously	.570	
28	Overall Passage Purpose	.564	
30	Writer-Reader	.552	
34	Channel of Presentation 2	.550	
17	Grammatical Resources 2	.550	
20	Weighting	.529	
32	Nature of Information	.512	
18	Grammatical Resources 3	.482	
5	Careful Local	.481	
25	Passage Length	.440	
27	Time Constraint	.425	
22	Knowledge of Criteria	.384	
4	Careful Global	.326	.303
26	Discernment 2	.302	
14	Lexical Resources 2		
23	Order of Items		
9	Expeditious Search 2		.656

10	Expeditious Search 3	.624
11	Expeditious Search 4	.603
15	Lexical Resources 3	.507
8	Expeditious Search 1	.504
2	Discernment1	.382
24	Channel of Presentation 1	.372
7	Rubric 1	.334
3	Grammatical Resources1	
13	Response Method1	
16	Functional Resource Imaginative	
6	Lexical Resources 1	
12	Functional Resource Ideational	

Extraction Method: Principal Component Analysis.

Rotation Method: Oblimin with Kaiser Normalization.

a. Rotation converged in 5 iterations.

However, as explained above, this can be reduced to two, namely Order of items and Functional resources. Functional resources represent high metacognitive skills which, in the context of this study, may not have been developed sufficiently by the students. Low loading for Order of items can be explained similarly that it is likely that most students simply completed the test in the order in which the questions were presented in the test paper. The remaining variables loaded significantly on either one of the components but a few of the variables have loaded significantly on both components. Careful global loaded on Component 1 at .326 and also on expeditious reading .303. However, the latter loading was only just above the threshold. This probably indicated that some students answered questions requiring expeditious reading but needed to use careful global instead. Rubric is represented on both but very significantly on Component 1. This is understandable as the rubric informs the reader about the type of reading that may be required in either case. One other result that requires some explanation is that Scanning

expeditiously has loaded quite strongly on Component 1 (.570), while all of the Expeditious search items (1, 2, 3, & 4) loaded on to expeditious reading. An explanation lies in the fact that expeditious search is at global level but scanning expeditiously is at local level. This would suggest that many of the students are still reading carefully to locate specific points of information. However, expeditious search skills, which are important for academic reading, were still underdeveloped.

The Component Correlation Matrix (Table 4.7) for a two-component model showed a much lower coefficient (.139) than for the model based on 6 components. This is evidence for an improved model based on two components.

Table 4.7 Component Correlation Matrix

Component	1	2
1	1.000	.139
2	.139	1.000

Extraction Method: Principal Component Analysis.

Rotation Method: Oblimin with Kaiser Normalization.

Thus, based on this empirical investigation, there is confidence that the scheme of variables used in this research is valid and that it is an appropriate scheme for answering the first research question which was to identify the cognitive processes utilised by students in the reading tests.

4.5 The variables

Having established through the correlation tests and factor analysis that only weak associations were found between these variables suggests that they were measuring different and distinct cognitive features and processes. This section now discusses the responses for each of these variables classified

under the component on which they had the highest loading. Each statement is presented with the particular cognitive feature being noted.

4.5.1 Variables with strong loadings on Component 1

Each variable is presented here in the order of the strength of their loading on Component 1 and the loading coefficient is included in brackets.

4.5.1.1 Discourse mode 2 (.628) and Discourse mode 1 (.625)

The passage presents the historical development of Zanzibar in an easy - to-understand way.

I think the way the paragraphs are arranged helped me to understand the passage.

Discourse mode 2 had the highest loading (.628) closely followed by Discourse Mode 1 (.625). Both variables had a high response rate (*M* = 3.08, *M* = 3.03, respectively) signifying a high degree of agreement with the statement. 81% for Discourse mode 2 (*n* = 163) and 77% for Discourse mode 1 (*n* = 156) of respondents were in overall agreement. It is clear that students' understanding is greatly aided by an understanding of historical genre in the case of Discourse mode 2 and by logical and chronological coherence in the case of Discourse mode 1 which they felt had greatly facilitated their comprehension.

4.5.1.2 Rubric (.613)

I found the test instructions easy to understand.

With a mean score of (*M* = 3.02), the majority of the students were in agreement or strong agreement with the statement (78%; *n* = 158). This indicates that the students perceived that the overall instructions as well as the individual questions were clear and easy to understand. This contrasts with the results for statement 7 where 64% agreed or strongly agreed that the rubric of that particular question was unclear. It is likely that the 22% (*n* =

44) who disagreed did so due to exam format familiarity or else guessed what was required.

4.5.1.3 Content knowledge (.607)

The article was mostly familiar to me due to my knowledge of the history of the connections between Zanzibar and Oman.

This variable with a coefficient of **(**.607) is closely related to Discourse mode 1 and 2 above. 75% (*n* = 152) showed overall agreement that familiarity with the historical connections between Zanzibar and Oman contributed to their comprehension of the passage. 25% (*n* = 50) were in overall disagreement that such content knowledge was of any use for comprehension. The mean score was high (*M* = 3.09).

4.5.1.4 Response method 2 (.592)

I found the variety of question types (e.g. true/false, short answer, gap filling, etc.) helpful in allowing me to show my skills.

This variable loaded strongly on Component 1 (.592) with a high mean score (*M* = 3.22) and with 81% (*n* = 163) in overall agreement with the statement. Only 19% (*n* = 39) were in disagreement. The evidence appears strong that the students responded well to the variety of question types as they felt that such variety allowed them to demonstrate their skills. It is possible that the students may have drawn on their many skills to respond to different types of response methods including self-management and prioritising. For those who disagreed, it is likely that they may not have applied self-management skills appropriately.

4.5.1.5 Overall passage purpose (.571)

Recognising the overall purpose of the passage helped my understanding.

The mean score on this statement was (*M* = 2.90) indicating a general agreement with the statement. 71% (*n* = 143) of the students believed that identifying the overall purpose of the passage aided comprehension.

However, 29% (*n* = 59) of the students appeared not to have recognised the importance of identifying the purpose and so were not using this process for comprehension. The loading coefficient was (.571).

4.5.1.6 Nature of information (.553)

I found most of the words on the passages were factual e.g. car, apple.

74% (*n* = 150) of students were in overall agreement that most of the words were concrete and factual. 26% (*n* = 52) disagreed or strongly disagreed with a mean score (*M* = 2.98) which is likely to reflect the fact that they found some abstract words challenging and still needed to augment their vocabulary.

4.5.1.7 Grammatical resources 2 (.542) and Grammatical resources 3 (.477)

I could easily understand 'it' referred to in line 7.

I found it easy to decide which of the three options best express the main idea of Paragraph 1 because the sentences were short.

The mean scores were (*M* = 2.96 and *M* = 2.76, respectively) indicating general agreement (74%; *n* = 149 and 62%; *n* = 104, respectively). These results present contrasting findings. In the first case (Grammatical resource 2), grammatical resources were correctly understood and applied (such as pronouns as in this statement) but that other grammatical resources such as syntax still required development (Grammatical resource 3).

In the latter case, despite the comments made by a number of theorists (see Chapter 2 Section 2.7.3.5) that shorter sentences can occasionally be more difficult to comprehend than longer sentences, the results here suggest that the short sentences actually aided the students' comprehension.

4.5.1.8 Writer-reader (.535)

This passage is suitable for Omani students.

On writer-reader relationship, there was a high level of agreement reflected in the mean (*M* = 3.08) with 76% (*n* = 154) in overall agreement. Nevertheless, 24% (*n* = 48) stated that they disagreed that the passage was suitable for Omanis. It is possible that the ones who agreed were able to draw on their prior knowledge and content knowledge.

4.5.1.9 Channel of presentation 2 (.520)

I was easily able to understand the passage even though there were no headings.

The mean score on this statement was (*M* = 2.83) with an overall agreement response of 67% (*n* = 136). This signifies that the absence of headings did not hinder the test takers' comprehension. However, one third of the responses were in disagreement (33%; *n* = 66), which seems to suggest that headings would have aided their comprehension. It is, therefore, necessary to investigate whether the inclusion of headings should be recommended or whether headings in this type of article (descriptive historical genre) would not aid comprehension but would fragment its inner unity.

In contrast, exactly the opposite response rates were elicited on statement 24 which is related to this statement as it deals with channel of presentation. However, the context is quite different as the provision of a time line might remove the necessity of reading the passage at all.

4.5.1.10 Scanning expeditiously (.519)

I was quickly able to find the information required to answer the questions in PART A.

The results showed overall agreement with the statement. 79% (*n* = 160) either agreed or strongly agreed and 21% (*n* = 42) disagreed or strongly disagreed with a mean score (*M* = 2.93). A plausible explanation is that the

majority of students were applying scanning expeditiously as intended and correctly answered the questions. In the Section (4.5.3) below, this supposition is tested using Kendall's tau test. The remaining students answered the question, probably by careful reading, which was not the cognitive skill that they were expected to have employed in this context (see Table 4.8 in Appendix 12).

4.5.1.11 Weighting (.504)

I think the number of marks given for each question was appropriate.

Regarding the fairness of the appropriation of marks per test item, 77% (*n* = 156) agreed that the various tasks were appropriately weighted with a mean score of (*M* = 2.98). Those who disagreed (23%; *n* = 46) were probably challenged in terms of time and decision-making.

4.5.1.12 Careful Local (.499)

I was able to answer question 9 by carefully reading the sentences in line 17 and 18.

There was a mean score of (*M* = 2.86) with 66% (*n* = 134) either agreeing or strongly agreeing. Thus the majority of the respondents answered the test question by appropriately applying careful reading at the local level. 34% (*n* = 68) of the respondents inappropriately applied some other process, or else did not need to use careful reading as the word in question was already part of their vocabulary (refer to Table 4.9 for other possible processes).

4.5.1.13 Knowledge of criteria (.477)

It was helpful to know how many marks were allocated to each item.

The mean score on statement 22 was high at (*M* = 3.11). There was 77% (*n* = 155) overall agreement that the knowledge of the criteria for scoring was helpful for the students. This is in contrast with 23% (*n* = 47) who disagreed

which indicated an underdevelopment of time management and decision making skills.

4.5.1.14 Passage length (.437)

I did not need to read every word of the passage in order to understand its meaning.

A high level of overall agreement (62%, *n*=125) with the statement indicated that the students were able to tackle a longer passage using the skills of skimming and scanning. However, more than a third (38%, *n* = 77) disagreed which indicated that they were using careful reading inappropriately. The mean score was (*M* = 3.01).

4.5.1.15 Careful global (.407)

I had to read other sentences carefully in addition to the sentence on line 16 in order to answer question 8.

These results show that 65% (*n* = 132) of respondents were applying careful reading at a global level; however, 35% (*n* = 70) did not use careful reading and have inappropriately used expeditious processes (e.g. skimming or search) or else simply guessed the answer to the test question but unwittingly disclosed that another process had been inappropriately applied. The mean score was (*M* = 2.82).

4.5.1.16 Time constraint (.400)

40 minutes was sufficient time to answer all the questions.

The mean score on this statement was (*M* = 2.72) with 63% (*n* = 127) of the students in overall agreement that the allocation of time was sufficient to answer the questions. However, over one third (37%; *n* = 75) disagreed which meant that they had not managed their time to best advantage.

4.5.1.17 Discernment 2 (.372)

I looked at the questions first before deciding whether to read the passage carefully or quickly.

With a mean score of (*M* = 3.03) it was clear that students were employing appropriate self-management skills, using discernment and devising a strategy. 73% (*n* = 148) of the students were in overall agreement regarding this strategy for tackling a longer passage. However, 27% (*n* = 54) still needed to develop discernment and self-management skills.

4.5.2 Variables with strong loadings on Component 2

4.5.2.1 Expeditious search 2 (.630)

I had to quickly search the whole passage to find the information required to answer question 11.

As expected, the results were similar to those for statement 8 with a mean of (*M* = 2.81) and 66% (*n* = 132) either agreeing or strongly agreeing with the statements. This indicates an appropriate application of expeditious search in reading. Those who disagreed (34%; *n* = 70) most likely applied careful reading which could have resulted in poor time management.

4.5.2.2 Expeditious search 3 (.622)

I had to quickly search the whole passage to find the information required to answer question 12.

On this question, the number in agreement is somewhat lower than for the previous two although still a majority (58%; *n* = 116). The question: (who moved the capital to Zanzibar?) could have been answered by some students by drawing on their prior knowledge of local history; thus the relatively lower percentage agreeing (42%; *n* = 86) does not necessarily imply that the others have used an inappropriate process. Another plausible explanation is that the four statements (8, 9, 10 and 11) were testing the same cognitive process which involved a search of the same text so that, in

answering the first two, some students may have noticed the information needed to answer the last two.

4.5.2.3 Expeditious search 4 (.595)

I had to quickly search the whole passage to find the information required to answer question 13.

Once again, the mean (*M* = 2.73) is somewhat lower than for statements 8 and 9. 63% (*n* = 127) agreed that it was necessary to search the whole passage to find the information but it is likely that at least some of the respondents who disagreed (37%; *n* = 75) may have already located the information in earlier searches for 8, 9 and 10.

4.5.2.4 Expeditious search 1 (.495)

I had to quickly search the whole passage to find the information required to answer question 10.

The mean score was (*M* = 2.93) with 72% (*n*=145) in agreement with the statement which implies that they utilised expeditious search reading in answering the question. The remaining 28% (*n* = 57) must have used a more careful process which would have been time consuming and could well have resulted in incorrect answers.

4.5.2.5 Lexical resources 3 (.482)

I needed to read beyond the sentence in line 7 in order to understand the meaning of 'the climate is tropical and humid'.

With a mean score (*M* = 2.66) and with 62% (*n* = 126) in overall agreement with the statement, these results indicate that, as in the previous question, the students were somewhat challenged in applying lexical resources. At least one of the two terms should have been known and this should have aided comprehension. However, 38% (*n* = 76) disagreed which suggests that

they either were able to infer the meaning or else were able to guess in a multiple choice context.

4.5.2.6 Discernment 1 (.386)

I found it difficult to decide whether to skim (fast read) or read carefully the whole passage in order to answer questions.

Table 4.8 shows that 65% (*n* = 130) either agreed or strongly agreed and 35% (*n* = 71) disagreed or strongly disagreed with the statement. The mean score was (*M* = 2.69). Those who agreed with the statement found it difficult to apply discernment and only the remaining respondents did not have much difficulty in applying the strategy. However, this does not suggest that most students did not apply discernment; rather it suggests that they found this strategy difficult to apply which is not surprising for such a high level metacognitive strategy for level four students.

4.5.2.7 Channel of presentation 1 (.348)

A time line of main events would have aided and supported my passage comprehension.

66% (*n*=133) of the students felt that a timeline would have helped their comprehension of a passage. The mean score was (*M* = 2.78). It is to be noted that results represent the students' preferences but nevertheless attention will be given to this issue in the discussion. 34% (*n* = 69) disagreed possibly because they were able to rely on textual features.

4.5.2.8 Rubric 1 (.315)

I was unsure whether long detailed answers or short answers were required in answering Part D.

Almost two thirds of the students (64%; *n* = 129) agreed that they were unsure about the level of detail required in answering while 36% (*n* = 73) did not have any difficulty in giving a short answer without detail. The mean

score was (*M* = 2.77). The rubrics in the test paper were not very clear as it simply stated ‘Answer the question’. Without specifying whether a short or more detailed answer was required.

4.5.3 Identifying the components: Descriptive analysis of participants’ responses and scores

Considering the variables which loaded high on each component, it appears that Component 2 is mainly expeditious reading. The four highest loadings were all types of expeditious reading either skimming or scanning (see Table 4.8 in Appendix 12). Thus, there is little difficulty in recognising Component 2 as being largely expeditious type of reading.

Having identified Component 2 as expeditious reading, it is tempting to immediately identify Component 1 as careful reading. However, a close examination of the high loadings on this component would not support this conclusion (see Table 4.9 in Appendix 13). While it is true that both careful global and local are among the loadings, they are by no means near the top of the table and it is noteworthy that scanning expeditiously actually loaded on to this component and loaded higher than either global or local careful reading. Keeping the research question in focus, what is being sought here is the actual processes the students used whether or not they fitted comfortably with what they might have theoretically been expected to do. Component 1, therefore, consists largely of self-management skills, drawing on familiarity with topic and with test format. Careful reading is involved in this component; for example, the highest loading was Discourse Mode 2, which most likely involved a careful reading based on cohesive devices. Likewise, Content Knowledge most likely involved careful reading as did Grammatical resources. However, loading on this component also was Scanning Expeditiously but in this particular case it was based on a text where prior knowledge or familiarity with the subject would have greatly aided comprehension. Knowledge of criteria, Time constraint, and Discernment also indicate that skills of self-management were being applied. All of these points are set out in Table 4.9 in Appendix 13.

Therefore, it cannot be easily concluded that Component 1 represents careful reading although it appears that there are elements of careful reading involved. Component 1 consists mainly of self-management skills and drawing on prior knowledge in order to comprehend what is being read which frequently did not require careful reading. Accordingly, Component 1 probably reflects some basic reading processes being used by students who have yet to develop reading skills drawing on the higher socio-cognitive skills which characterise academic reading.

Finally, the scores of students on test questions were examined to test whether choosing the intended process corresponded with a significantly higher test score. These are presented in Table 4.10:

Table 4.10 Summary of strong loading factors under Components 1 and 2

	Variables	Coefficiency result from Component Matrix	Test item corresponds to the component	Total score of 202 students per test item	Students response to the statement in the questionnaire corresponding to the test item	
					Agreement %	Disagreement %
Component 1	Scanning expeditiously	.519	1	189	79 %	21 %
			2	168		
			3	172		
			4	104		
	Careful local	.499	9	77	66 %	34 %
	Grammatical resources 3	.477	6	246 (2 marks each question)	62%	38%
	Careful global	.407	8	109	65%	35%

	Variables	Coefficiency result from Component Matrix	Test item corresponds to the component	Total score of 202 students per test item	Students response to the statement in the questionnaire corresponding to the test item	
					Agreement %	Disagreement %
Component 2	Expeditious search 2	.630	11	125	66 %	34 %
	Expeditious search 3	.622	12	140	58 %	48 %
	Expeditious search 4	.595	13	108	63 %	37 %
	Expeditious search 1	.495	10	140	72 %	28 %
	Discernment1	.386 Part B: the questions also examine expeditious skimming Q 6 examines also grammatical resources 1	5	292 (2 marks each question)	64 %	36 %
			6	246 (2 marks each question)	57 %	43 %
	Rubric 1	.315 This also covers expeditious search above	10	140	72 %	28 %
			11	125	66 %	34 %
			12	140	58 %	48 %
			13	108	63 %	37 %

Only those variables that were directly represented in a test item/question were included in this table. The variables not represented here mostly related to the test passage rather than answering a test question (see examples under column 'Where in test paper (Item/question?)' in Table 4.8 and Table 4.9 in Appendices 12 & 13). Each component is now examined in greater detail.

4.5.3.1 Component 1

The Kendall's tau test was run on the total level of agreement in the questionnaire responses and the total scores of related test questions for Component 1 (see Table 4.11):

Table 4.11 Kendall's tau test of total questionnaire scores and total test items scores for Component 1

			Test Scores Component 1	Questionnaire Scores Component 1
Kendall's tau_b	Test Scores Component 1	Correlation Coefficient		.020
		Sig. (2-tailed)		.699
		N		202
	Questionnaire Scores Component 1	Correlation Coefficient	.020	
		Sig. (2-tailed)	.699	
		N	202	

The results of the test were (T= .020, n = 202, p = .699). The significance level was well outside of the range for significance (p < .05). Thus, there was insufficient evidence for rejecting the null hypothesis. Accordingly, on Component 1 taken as a whole, there is insufficient evidence of an association between answering the test questions correctly and using the intended process. In order to understand this result, it was decided to examine each statement separately running the appropriate tests on each.

- ***Scanning expeditiously (Statements 1 against test items 1, 2, 3, 4)***

The statement on Scanning expeditiously was represented by four test items. There was 79% agreement indicating that the majority of students were scanning expeditiously. As can be seen, all of the test scores were over 50% ranging from question 1 where 93.5% of the students answered correctly to 54% answered correctly on the fourth question. The mean questionnaire scores for those who answered each item correctly and incorrectly are presented in Table 4.12 below:

Table 4.12 Mean scores of statement for test items 1, 2, 3 and 4

	Questionnaire mean scores for test items answered correctly	Questionnaire mean scores for test items answered incorrectly
Item 1	2.9409	2.8125
Item 2	2.9353	2.9063
Item 3	2.9697	2.75676
Item 4	3.03061	2.83654

The table shows that in all four test items, those who answered correctly had a higher mean questionnaire score than those who answered incorrectly,

albeit that these differences do not appear to be great. Thus, it would appear, prima facie, that the students who chose the appropriate process also managed to a large extent to answer the question correctly. However, running the Kendall's tau test on the total scores for the four questions led to a more subtle analysis of the situation (see Table 4.13).

Table 4.13 Kendall's tau test of total scores of statement 1 against corresponding total test item scores

Correlations

			Items 1, 2, 3, 4 Scores for statement 1 Com1
Kendall's tau_b	Questionnaire Score Statement 1 Component 1	Correlation Coefficient	.115
		Sig. (2-tailed)	.066202

As can be seen from the Kendall's tau result, there is not sufficient evidence to reject the null hypothesis of 0 difference and the correlation coefficient showed only a weak association (T = .115, p = .066). However, the probability returned was close to the threshold (p < .05) and it was decided to run a series of Mann Whitney U tests for each item but none of these proved to be significant (see Tables 4.14, 4.15, 4.16 and 4.17 in Appendix 14). Accordingly, the hypothesis of no significant difference is accepted. There is, thus, no evidence in the case of questions 1, 2, 3 and 4 of a significant association between using the intended process and answering the questions correctly. Consequently, the validity of test items 1, 2, 3 and 4 is in doubt as these items did not allow for a discrimination to be made between those who used the intended cognitive process and those who answered using an alternative process.

- ***Careful local (Statement 5 against test item 9)***

Under the question of Careful local (Statement 5), while 65% of the students were in agreement that they were able to answer question 9 by reading carefully, only 38% answered the question correctly which indicates that even though they might have applied careful reading, they did so without being able to comprehend the meaning of what they were reading. Of the students who answered the question correctly their mean questionnaire score was 2.85 with a mean score of 2.88 for those who answered incorrectly which, prima facie, suggests that choice of process had little effect on the outcome of the test item. This is confirmed by the result of the Mann-Whitney U test which returned a z value of -.212 and a significance of p = .832 as shown in Table 4.18.

Table 4.18 Mann-Whitney U test for questionnaire statement 5 between students who answered test item 9 correct and incorrect

Test Statistics[a]

	Questionnaire Score Statement 5 Component 1
Mann-Whitney U	4776.500
Z	-.212
Asymp. Sig. (2-tailed)	.832

a. Grouping Variable: Item 9 Scores for statement 5 Com 1

Since the p value is > .05, there is no statistically significant difference (Pallant, 2010) between the test takers who answered correctly and those who answered incorrectly based on their chosen process (Statement 5). Despite agreeing with the statement (and by implication acknowledging that they were using the intended process), this was not reflected in a significantly higher number of the students giving the correct answer. Thus the validity of item 5 is in doubt.

- ***Grammatical resources 3 (Statement 18 against test item 6)***

62% (246 marks; 2 marks each test question) were in agreement with the statement (Grammatical resources 3) that they were able to choose the correct option due to the sentences being short and 61% (246 marks out of 404) of the students answered correctly. The mean questionnaire score for those who answered correctly was 2.84 compared with a mean score of 2.64 for those who answered the test question incorrectly. So it appears that those who used the intended process also tended to give the correct answer. However, this is not supported by the Mann-Whitney U test which returned a z value of -1.139 and a significance of *p*= .255 (see Table 4.19).

Table 4.19 Mann-Whitney U test for questionnaire statement 18 between students who answered correct and incorrect test item 6

Test Statistics[a]

	Questionnaire Score Statement 18 Component 1
Mann-Whitney U	4299.000
Z	-1.139
Asymp. Sig. (2-tailed)	.255

a. Grouping Variable: Item 6 Scores for statement 18 Com 1

Thus, there is no evidence of a significant difference between the test takers who answered item 9 correctly and those who answered incorrectly in terms of questionnaire statement 5 scores. Accordingly, the null hypothesis is retained. Thus, it is concluded that choosing the intended process was not significantly associated with giving the correct answer and conversely that those who chose some other process were not significantly more likely to obtain an incorrect answer. Thus, the validity of test item 6 is in doubt.

- ***Careful global (Statement 4 against test item 8)***

With regard to Careful global, 65% of the students agreed with the statement which indicated that they were applying careful global reading and 54% (109 marks out of 202) managed to answer test item 8 correctly. This contrasts with the very low score on Careful local but even with 54% (77 marks out of 202) on this one raises the issue of whether many of the students were actually using careful reading. The mean questionnaire score was 2.90 for those who answered the test item correctly and 2.73 for those who answered incorrectly. The Mann Whitney U test was run and indicated that there was no significant difference (z = -.650, p = .516) between using the intended process and getting a correct answer on the test question (see Table 4.20). Once again, the validity of this test item is in doubt.

Table 4.20 Mann-Whitney U test for questionnaire statement 4 between students who answered correct and incorrect test item 8

Test Statistics[a]

	Questionnaire Score Statement 4 Component 1
Mann-Whitney U	4811.000
Z	-.650
Asymp. Sig. (2-tailed)	.516

a. Grouping Variable: Item 8 Scores for statement 4 Com 1

In summary, The Kendall's tau test found no significant difference on Component 1 taken as a whole and answering test questions correctly. Mann Whitney U tests based on individual questionnaire items and corresponding test items did not provide evidence of any significant differences for any association between using the intended process and answering the related test question(s) correctly. This will be commented on in detail in the discussion chapter.

4.5.3.2 Component 2

In the case of Component 2, the Kendall's tau test was run on the total scores for the questionnaire responses and the total scores for the test questions. The results are shown in Table 4.21:

TABLE 4.21 Kendall's tau test of total questionnaire scores and total test items scores for Component 2

Correlations

			Questionnaire Scores
Kendall's tau_b	Test Scores	Correlation Coefficient	.183**
		Sig. (2-tailed)	.001
		N	202

**. Correlation is significant at the 0.01 level (2-tailed).

The test returned a moderate level of association (T= .183) but, nevertheless, a significant result (p = .001). Thus, there is very strong evidence of an association, albeit a weak one, between using the intended process and answering the test questions correctly. This result is in contrast with what obtained for Component 1. Component 2 appears to be expeditious reading and the Kendall's tau test result confirms that students who applied expeditious reading appropriately were also more likely to answer the questions correctly.

In conclusion, it appears that where expeditious reading is required that the students who used this process tended to answer the test questions correctly and conversely those who did not apply expeditious reading tended to answer the test questions incorrectly.

Questions 10 to 13 were concerned with testing expeditious search and the rates of agreement with the statements were high indicating that the appropriate process was used. Generally, the test scores show that over 50% (Q11: 125 marks, Q12: 140 marks, Q13: 108 marks, Q10: 140 marks, respectively) of the students scored correctly and this corresponds with the high rate of agreement with the intended process. Two test questions, 5 and 6 represented Discernment 1 and there was fairly a strong agreement with the statements (64% and 57%, respectively). The test scores were well above 50% indicating that it was most likely that those who chose the appropriate process also answered the questions correctly. However, the statement related to the degree of difficulty in applying discernment but the test scores confirm that, despite finding it difficult to decide which process to use, they nevertheless managed to answer the questions correctly. The final variable which was Rubric 1 concerned uncertainty about the level of detail required to answer the questions and the table shows there was a high level of agreement that the rubric was not clear. The absence of a clear instruction regarding the level of detail required in the answer may raise issues about the fairness of these four questions. However, in every case test item scores were high (none below 50%). Examination of the test scripts revealed long answers where only one sentence was required and this may have presented difficulties for the test markers in deciding whether to award a score or not.

4.6 Summary

This chapter has presented the results and findings of the data that was collected to answer the first research question:

“What are the cognitive processes by which students engage with the texts and tasks in reading tests?”

A natural experiment was conducted in the form of a simulated test using a questionnaire (see Chapter 3 Section 3.4). The data was represented by overall descriptive statistics and inferences were made using SPSS. Correlations tests were conducted which established that the variables were

measuring distinct cognitive features. Factor analysis further supported the robustness of Khalifa and Weir's (2009) model in that no strong associations were found which might have cast doubt on the robustness of the model. Factor analysis led to a two component solution to represent the data. However, while expeditious reading was one of the components in the factor analysis, the other component could not be simply equated with careful reading but included other processes.

Running Kendall's tau test on Component 2, by testing test results against questionnaire responses, returned a significant level for an association. Thus, for the items on Component 2, it was concluded that the students who chose the intended process also tended to score correctly on the test items. Therefore, the validity of test items measuring expeditious reading is supported. However, in the case of Component 1, the validity of the test items is in question as only one association was found.

Consequently, there is confidence in the validity of inferences made on the basis of the cognitive processes in Component 2 (expeditious reading) by which students engaged with the texts and tasks in the reading tests. In contrast, the validity of inferences on the basis of the cognitive processes in Component 1 cannot be established with confidence. However, it is acknowledged that there is a limitation to these findings as they were based on a single version of a test using one questionnaire and a relatively small sample size of 202 students.

In the following chapter, these findings are discussed in the light of the relevant literature.

Chapter 5 Discussion I: Verbal Protocol Analysis

5.1 Introduction

The purpose of this chapter is to critically discuss the findings from the verbal protocol analysis (Chapter 4) which was conducted in order to answer the first research question:

"What are the cognitive processes by which students engage with the texts and tasks in the reading tests?"

Weir et al. (2009) drew attention to the importance of test tasks and reading activities in assessment. This is so in consideration of the inferences which are intended to be drawn from such tasks and activities about students' abilities to perform in the real world using a second language. For this reason, they argue that tasks and activities should be considered in relation both to the cognitive processes and the contextual features which are involved in their completion. This chapter focuses on the cognitive processes based on the 34 variables which were derived from the literature and discussed further in the methodology chapter (see Chapter 2 Section 2.7 and Chapter 3.4). The contextual features are analysed and discussed in Chapters 6 and 7.

The 34 variables (see Chapter 4 Section 4.1) were first tested for any association or overlap using correlation (see Chapter 4 Section 4.3). This provided significant evidence that each of the variables were measuring distinct cognitive processes (see Chapter 4 Section 4.3 and Section 4.4). Factor analysis established that the cognitive processes could be categorised under two components (see Chapter 4 Section 4.4). Component 1 involved a number of *basic readings processes* (e.g. self-management and drawing on prior knowledge). In contrast, Component 2 was easily identified as *expeditious reading*.

Student responses on variables which had strong loadings on Components 1 and 2 (shown on Chapter 4 Table 4.5 and Table 4.6) were compared with the students' test scores. This was done to determine whether there was any

evidence that choosing the intended cognitive process resulted in a significantly higher likelihood of answering the test questions correctly. Additionally, this would also determine the extent to which choosing an alternative process to the intended one increased the likelihood of giving an incorrect answer in the test (see Chapter 4 Section 4.5.3).

Both of these components are now discussed in turn with reference to the relevant literature.

5.2 Component 1: Basic reading processes

Component 1 comprised the following variables: *Scanning expeditiously, Careful local,* the knowledge and use of *Grammatical resources 3 and Careful reading at global level* (see Chapter 4 Table 4.10), which were found to load significantly. When the students' responses to the statements were tested against the exam scores, the results showed that the choice of cognitive process on each of these variables did not have any significant effect on answering the questions correctly (see Chapter 4 Section 4.5.3.1). Each of these variables are now examined and discussed.

5.2.1 Scanning expeditiously

The analysis showed that there was no evidence that choosing the process of scanning expeditiously in the questionnaire had any significant effect on answering the test questions correctly (see Chapter 4 Section 4.5.3.1 Component 1). This result is of some concern particularly as a number of authors, (e.g. Khalifa and Weir, 2009; Weir, 2013) have emphasized the importance of this skill especially for students progressing to undergraduate level and who would be faced with extensive reading lists within a few months. As was discussed in (Chapter 2, Section 2.2), this skill has been somewhat neglected in teaching and testing in the main developed countries, i.e. US and UK (Weir, 2013). However, the response rate in the Omani test situation showed a high level of agreement with *Statement 1: I was quickly able to find the information required to answer the questions in PART A.*

Nevertheless, agreeing with this process did not significantly affect their answering the test questions correctly:

PART A: Circle T if the statement is true, F if it is false. (4 marks)

1)	Pemba is an island to the north of Unguja.	T	F
2)	The climate in Zanzibar is mild and dry.	T	F
3)	Zanzibar was called the Spice Islands	T	F
4)	Tourism dominates the Zanzibar economy.	T	F

In fact, the students who chose an alternative process were just as likely to answer correctly. This raises the issue of the validity of test items 1, 2, 3 and 4 above in testing expeditious scanning. The alternative processes that the test takers might have used include careful reading. Of course using careful reading would still lead to answering the questions correctly. Since there is no control of the time the test takers can spend on this question (there is a time limit for the overall test), it is, therefore, possible for the students to use careful reading spending more time and using other cognitive processes in an exercise for which this was not intended. On this very point, Alderson (2000) has drawn attention to the need for comprehension in this context to be strictly time bound reflecting real world or academic scenarios. This is a requirement for authenticity (see Chapter 2 Section 2.7).

A possible explanation for test takers spending more time than was intended has been revealed in studies of L1-L2 transfer. These have commented on the fact that Arabic is a highly inflected language with high salience of features which require great attention to individual words rather than the meaning of a group of words (Palmer et al., 2007). This is a plausible explanation for the higher response rate based on a process of local expeditious reading rather than global (Randall, 2009).

Moreover, comprehension can be affected by slow word recognition as underlined by many authors (e.g. Adams, 1990, 2004; Samuels, 2004) which

in turn may affect the time and type of reading employed whether expeditious or careful. In the case of expeditious reading, where reading is time-constrained it seems likely that less proficient second language readers may not have had sufficient practice or may not have developed rapid word recognition. Accordingly, items designed to test these types of expeditious skills may be inappropriate for unpractised students (see Chapter 2 Section 2.3) and may produce results whose validity may be in doubt. Although expeditious reading is widely recommended as a strategy for passing some second language proficiency tests, it is not well defined in the literature. Furthermore, it is therefore also possible that expeditious reading is not a valid construct and doubts must remain about whether it can be measured.

The effect of lack of rapid word recognition on reading and comprehension for EFL learners is also acknowledged by, for example Yoshimura (2000) and Koda (2004). In fact, Koda (2004) has identified this problem in terms of "restricted word recognition skills" which would result in time and energy being expended on visual fixation on individual words rather than drawing on multiple sources of information. The test is time-bound for the very reason that students should not be expending so much time on word recognition but should be able to infer the meaning from the overall context. Instead, students in this situation would be less likely to engage in hypothesising and predicting meaning than more proficient L2 readers (Koda, 2004). An earlier alternative explanation is that they may have drawn on their prior background knowledge rather than using more abstract words which is a characteristic of more advanced L2 readers (Coady, 1979).

Another plausible explanation is to be found in the difficulty of constructing types of questions which are designed for a selection between true or false or yes/no responses (see Chapter 2 Section 2.7.2.1). These four questions are ill constructed; for example, consider the structure of test item 2: *The climate in Zanzibar is mild and dry*. It is possible for a test taker to answer this question using their own judgment to decide the veracity of the statement but this is not what is intended by the test designer. What is intended is that the test taker goes back to the text and discovers what claim is made in the passage (Khalifa and Weir, 2005, discussed in Chapter 2

Section 2.7.2.1). In other words, it is intended that the test taker understands what the article is stating about the climate, rather than relying on logic and problem solving as Valette (1977) has cautioned. Despite the benefits of test items of this kind for large scale testing (such as in the context of this study), in this particular test paper, true/false type questions have not really assisted in discriminating between strong and weak students, which is the goal of such exercises as was pointed out in (e.g. Hughes, 2003; Khalifa and Weir, 2009; Weir, 2013, see Chapter 2 Section 2.7.2.1).

Regarding the order of items in the design of these types of test items, Khalifa and Weir (2009) have drawn attention to the need for randomisation in test items for validity. In this particular test, the test questions followed the logical structure of the original passage. In doing so, the test designers have defeated the object of the exercise which was to test students' ability to scan the whole text. As the test items were presented here, once the first item was found, the remaining items could be quickly found without scanning.

Thus, it is not possible to draw inferences on these four questions relating to the test takers' ability to scan expeditiously as intended in such an exercise. Indeed, it may be that it is simply inappropriate to test L2 students on skills which are, as yet, underdeveloped and insufficiently practised. Furthermore, due to the paucity of theoretical foundation for expeditious reading, it remains a possibility that the construct itself may lack validity. Consequently, there is a need for caution when drawing inferences based on expeditious reading.

5.2.2 Careful local

Reading at careful local level involves "establishing accurate comprehension of explicitly stated main idea within a sentence" (Weir et al., 2009, p.101). Lexis and syntax have an important role to play at this local level. The test item related to this feature was:

PART C: Circle the est answer: (2 marks)

9) Find the word *prosperous* in line 17. Choose the best meaning for this word in the context.

a) wealthy b) poor c) dangerous

The analysis for this test item showed that 65% of the test takers had chosen the intended process (careful reading local) and the remaining 35% chose some other process in response to Statement 5*: I was able to answer question 9 carefully reading the sentences in line 17 and 18:*

(Line 17 & 18):
Because of the prevalence of the spice trade, Zanzibar was known as the Spice Islands. Rich and *prosperous*, Zanzibar caught the attention of other countries.

However, the mean score for both groups (those who used the intended process and those who used some other process) was found to be very close and the results of the Mann Whitney U test revealed no significant difference between the mean scores of the two groups (Chapter 4 Section 4.5.3.1). In other words, choosing the intended process did not result in a significantly higher mean score than choosing an alternative process. Accordingly, this test question fails to discriminate between those who chose the intended process and those who did not, and therefore the validity of this question is in doubt and, consequently, valid inferences could not be made regarding the students' readiness for academic reading using the process of reading carefully at local level.

It is unlikely that those who used an alternative process used reading at careful global level (e.g. by seeking to comprehend by means of information across sentences), as the rubric clearly directs that students to line 17 which

points them to careful reading locally. Nor could expeditious scanning have helped as they would still need to understand the word “prosperous” in its context. It is therefore likely that these test takers either guessed the meaning in a three way multiple choice situation or else the word “prosperous” was already part of their vocabulary and they were then able to directly answer the question without reference to the text.

A plausible explanation for answering correctly using a different process is provided by Stanovich (1980) who points out that some weaker readers may have inferred the meanings of the words through the greater sensitivity which poor readers have to contextual constraints. Stanovich (1980) claimed that good readers were often less sensitive to contextual effects than weaker readers (Stanovich (1980) cited in Samuel and Kamil, 1988, pp.32-33). Cooper (1984) has commented on the challenge faced by second language college students whose pre-university education had been conducted only through the medium of their native language, as is the case with the respondents in this study. In secondary education in Oman, English is taught discretely as a second language but other subjects are taught through the medium of Arabic. By comparison, Cooper (1984) found that students, whose pre-university education was through the intended second language, had a much greater vocabulary level and competence. Thus, it seems likely that the majority of the Omani students in this study were impeded in their comprehension due to their vocabulary deficit and also to their poor understanding of syntax. Comprehension in this test item hinges on appreciating the role of the word “and” (line 17 mentioned above) in linking two adjectives “rich and prosperous” in a way that implies that the second adjective ‘prosperous’ is an extension of the first adjective ‘rich’.

This could explain why there was no significant difference between the test scores of the two groups based on the process used, because those who chose the intended process were unable to apply it effectively due to their lexical deficit and were, consequently, no more successful in answering correctly than the group which used an alternative process. Indeed, this is confirmed by the low loading for lexical resources on this component which was less than .3 (see Chapter 4 Section 4.4).

It is likely however, that some of the test takers used guesswork given that multiple-choice options were available. However, the questionnaire findings reveal that the majority of Omani students were applying careful reading at local level but without effect. Those who answered correctly may still not have comprehended correctly due to the difficulty in constructing effective MCQs (Hughes, 2003, see Section 2.1 above). Hughes (2003) has pointed out the many difficulties that arise in constructing multiple choice questions include the use of ineffective distractors. In this case, three options are provided for answering. The first two 'wealth' and 'poor' are antonyms and third option 'dangerous' is unrelated in the context. So logically, without comprehension, the students could eliminate the third distractor increasing their chances of answering by guessing to 50% without reference to the text.

5.2.3 Grammatical resources 3

This linguistic demand was one of a set of test questions which measured comprehension based on the knowledge and use of grammar (see Chapter 2 Section 2.7.3.5). It was already found from the analysis (see Chapter 4 Section 4.5.3.1) that choosing the intended process did not significantly increase the likelihood of the students getting the correct answer and, the corollary of this was that choosing an alternative process did not significantly decrease the students' likelihood of answering correctly. This test question did not discriminate based on the choice of the process and therefore, its validity is in doubt. Again, one plausible explanation centres on multiple choice questions and difficulties surrounding their construction for validity as discussed in previous sections (2.1 and 2.2).

Statement 18 was designed to discover whether the test takers comprehended by means of grammatical resources:

> *Statement 18: I found it easy to decide which of the three options best express the main idea of Paragraph 1 because the sentences were short.*

In the exam this was tested by question 5:

> **PART B: Circle the best answer: (6 marks)**
>
> **5)** What is the main idea of **paragraph 1?**
>
> a) Zanzibar is an island nation off the coast of Africa.
>
> b) Zanzibar is a historically important and interesting place.
>
> c) Zanzibar is full of tourists who want to see the clove trade.

62% were in agreement on being able to find the main idea in Paragraph 1 where, not only were the sentences short but the paragraph itself was short. According to Shiotsu (2003) and Shiotsu and Weir (2007), syntax as opposed to lexis should play a central role in comprehension (Chapter 2 Section 2.7.3.5). In other words, syntactical parsing should aid the test takers in overcoming lexical deficit. However, the Mann Whitney U test found no significant difference in using a process based on the application of grammatical resources by drawing on syntax over lexis. Those who used an alternative process were no less likely to complete the task correctly (Chapter 4). This is evidence that, in this test, the test takers were not able to draw on syntactical resources.

The findings of this study suggest that the syntactical structure of the sentences did not significantly enable the students to overcome their lexical challenges through using the available syntactical resources. Drawing on recent research, Hall and Durán (2009) showed that more proficient readers were able to avoid reliance on lexical resources that were based on L1 and were able to accurately infer meaning by 'direct semantic mapping' (Hall and Durán, 2009, p.27). In view of the non-significant results of the Mann Whitney U test, it is clear that most of the test takers were unable to draw on grammatical resources to overcome lexical deficit in comprehending the passage. The results of this study indicate that most of the test takers were probably relying on comprehension by means of reference to their L1 and were unable to 'think in English.' The differences in opacity between Arabic

and English have already been discussed in Chapter 2 Section 2.7.3.5. There, it was explained how Arabic is less opaque in phoneme-grapheme compared with English which is more opaque in both. More proficient L2 readers can overcome this difficulty by being able to draw on syntactical resources or by direct semantic mapping. While it is expected that test takers at level four (the highest level in the Foundation Program, see Chapter 1 Section 1.3) would be able to draw on grammatical resources, it is clear in this study that most of them have not yet developed this capacity. This confirms the earlier conclusions reached in the light of Stanovich (1980) (Section 2.2 above) who claimed that less proficient L2 readers were more sensitive to contextual features than more proficient readers. Nevertheless, Hall and Durán (2009) have drawn attention to the fact that, where idiosyncratic lexicogrammatical features of L2 closely resembled those of L1, as is the case with the Arabic term 'و' which means 'and' in Arabic, that comprehension was aided and this may partially explain why those test takers who used an alternative process still obtained scores similar to those who used the intended process .

However, the conclusion is that the test question intending to measure grammatical resources lacks validity based on the result of the Mann Whitney U test.

5.2.4 Careful global

The analysis of the results for careful reading at global level showed that there was little difference in answering the test questions correctly between those who used the intended process and those who used an alternative process. The Mann Whitney U test results showed that there was no significant difference between the two groups and this calls into question the validity of the test item in measuring careful reading global. Statement 4 of the questionnaire measured the extent to which test takers used careful global for comprehension:

I had to read other sentences carefully in addition to the sentence on line 16 in order to answer question 8.

This was tested in the exam by question 8:

PART C: Circle the best answer:(2 marks)

8) Find the word *extensively* in line 16. Choose the best meaning for this word in the context.

a) not very much b) a lot c) occasionally

The questionnaire results revealed that 65% of the students used careful reading global and 35% used a different process than that intended (see Chapter 4 Section 4.5.1). However, there was no significant difference in the test scores obtained by either group and this casts doubt on the validity of the test item as a measure of careful reading global (see Chapter 4 Section 4.5.3.1).

The rubric of question 8 directs the test takers to line 16 where the word 'extensively' occurs, inferring that the meaning of the word (if it was not already known) required "making propositional inferences" based on a careful global reading (Khalifa and Weirs, 2009, p.101). In other words, to infer the meaning of this word required a careful reading beyond the sentence in which it occurred. The choices available to the test takers could have encouraged some test takers to guess incorrectly.

Those who used the intended process did not answer significantly more successfully than who used an alternative process. This means that although 54% of those who used the intended process answered correctly, this was not significantly higher than those who used an alternative process. 46% of those who used the intended process still got an incorrect answer and this demonstrates that a significant number of the test takers did not use careful global reading effectively. It has already been pointed out, in relation to

careful reading local, that a significant number of test takers had difficulty with that process. A fortiori, they could hardly be expected to be more effective using careful global given the difficulties they had with understanding syntax and lexis at local level. In relation to those who used a different process than the one intended, these test takers were not significantly less successful than those who used the intended process. It is possible that these test takers either used logic or guesswork or, it is also possible that some of them already had the word ‘extensively’ in their vocabulary.

In this context, Bernhardt (1991) cited in Randall (2009, p.119), observing eye fixation of L2 readers compared with L1 readers, has noted that those with low syntax knowledge tended to fixate more on functional words than L1 readers. For L1 readers, eye fixation tended to be on content words but L2 students with weak syntactical knowledge needed to pay much more attention to functional words such as the adverb ‘extensively’ in this study. The more proficient readers would have been able to infer the meaning of the adverb ‘globally’ from the previous sentence which commented on “Zanzibar trading widely with many other countries” in the test passage. It is possible that the less proficient readers were more fixated on the functional words within the sentence. If this is the case, then it means that a significant number of those who used an unintended process were relying on reading at careful local level instead of at global level. Therefore, they would have been expected to score significantly less than the test takers who used the intended process. The fact that there is no significant difference based on test scores strongly suggests that this test item lacks validity as a measure of careful global reading. The issue is that test item 8 fails to discriminate between those using the intended process and those using an alternative process. It is likely that the distractors in the multiple choice questions failed to allow for this discrimination to occur. A closer examination of the three distractors available to answer the questions shows that one of them ‘occasionally’ is an adverb form just as the word ‘extensively’ is. This is a little unfair as that connection could mislead the test taker deliberately and this element of bias alone would render the test item invalid. The remaining two options did not represent single worded adverbs:

‘not very much’ and ‘a lot’. Using alternatives such as ‘broadly’ and ‘narrowly’ instead would have at least removed the element of bias from the distractors. Hughes (2003) draws attention to clues in the options which affect validity. In this case, the clue in the options (adverb form to adverb form) could have been unfairly misleading for some test takers. Consequently, inferences regarding these test takers ability to apply the careful global reading in an academic situation would also lack validity.

In summary, the Kendall tau’s test found no significant differences on features categorised under Component 1 taken as a whole and correctly answering test questions (see Chapter 4 Section 4.5.3.1). However, the complexity of reading has been much commented on by a number of authors (e.g. Alderson and Urquhart, 1984; Perfetti, 1985), so that it is not possible to be definitive about what alternative processes students may have used instead of the intended processes. Grabe (2009) mentioned a range of processes, including the use of prior knowledge, to compensate for their deficits in other linguistic processes such as word recognition, syntactic parsing and meaning formation. Grabe continued by asserting that “We do this seemingly without effort and with all processes synchronizing in time” (Grabe, 2009, p.14). Thus, the use of alternative processes by some Omani students without any deleterious effects in answering these test questions, confirms the view that Component 1 comprised a number of basic processes such as self-management and perhaps even guesswork (see Chapter 4 Section 4.4).

5.3 Component 2: Expeditious reading

The analysis found a significant result for the items loading on Component 2 and, although the association was a moderate one, the significance level was high (p = .001) (see Chapter 4 Section 4.5.3.1). This result confirms that test takers who applied expeditious reading appropriately were also more likely to obtain the correct answer to the test items. The features strongly loading on Component 2 included *Expeditious search 1, 2, 3 and 4*, *Discernment 1* and *Rubric*.

These features are discussed in the following sections. Other features of expeditious reading also loaded on Component 2 but below the 0.3 threshold and yet other features of expeditious reading were not represented by test items (see Chapter 4 Section 4.4).

5.3.1 Expeditious search 1, 2, 3 and 4

These were tested by test items 10 to 13:

PART D: Answer the questions. **(4 marks)**

10) What protects the beaches?

_____

11) What did Zanzibar become the largest producer of?

_____

12) Who moved his capital to Zanzibar?

_____

13 What is the origin of most of the population in Zanzibar?

_____

High rates of agreement in the questionnaire indicated that the intended process was being used by most of the test takers. It was found that test scores relating to these items were all over 50% (Q11= 62%, Q12= 68%, Q13= 53%, Q10= 68%, respectively). This showed that those who selected the intended process, *expeditious reading*, also tended to score significantly higher than those choosing an alternative process.

The statements relating to these four questions all revealed high levels of agreement by the test takers indicating that they were selecting the intended processes. The Kendall's tau test results confirm that the test takers who used the intended process were also significantly more likely to answer the questions correctly (see Chapter 4 Section 4.5.3.2). The result also confirms that those test takers who used an alternative process were significantly more likely to return an incorrect answer. This is strong evidence to support

the validity of these test items and that, consequently, there is confidence in inferences drawn about the test taker's ability to apply expeditious search reading process at academic level. This is supported in Khalifa and Weir (2009) who claimed that in the case of short answer type questions there was greater certainty that the result was due to comprehension rather than to any other factor (see Chapter 2 Section 2.7.2.1). Admittedly, there can be a problem with short answer type questions in that it requires writing where reading is the skill being tested. The results in the test, however, suggest that writing short answers did not greatly affect the outcome and that these test items enabled a valid discrimination to be made between the results obtained by those who used the intended process and the significantly lower scores by those who used an alternative process. This is convincing evidence that the Omani students are reading selectively and efficiently and thus, should be able to effectively transfer these skills to a more academic scenario. This is a most auspicious result in the Omani context given that the authors have commented on the often underdeveloped nature of the skills of expeditious reading in some developed Western countries where English is the first language. For example, Weir (2013) has already commented on the tendency for developed L1 countries to emphasise the importance of careful reading over expeditious and that the latter is often underdeveloped in these contexts (see Chapter 2 Section 2.2). In fact, Weir (2013) goes even further by claiming that, despite its inclusion in the CEFR and its importance for academic reading, it has not been explicitly tested in many high-stakes examinations in the UK (see Chapter 2 Section 2.2). However, it is readily admitted that, in the case of this test, the passage was relatively short and also there was only one passage to read. A more robust test of expeditious reading might be possible with a longer passage and at least one other passage (see Chapter 2 Section 2.7.2.6) with perhaps a slightly longer time allocation.

5.3.2 Discernment

Discernment 1 was represented by two test questions (5 & 6):

> **PART B: Circle the best answer: (6 marks)**
>
> **5)** What is the main idea of **paragraph 1?**
>
> a) Zanzibar is an island nation off the coast of Africa.
>
> b) Zanzibar is a historically important and interesting place.
>
> c) Zanzibar is full of tourists who want to see the clove trade.
>
> **6)** What is the main idea of **paragraph 3**?
>
> a) Many things have happened in Zanzibar's long history
>
> b) Little is known about the early history of the islands.
>
> c) The ivory trade was important in the past.

These questions represented variables for measuring discernment. Responses in the questionnaire to Statement 2 measured the extent to which discernment was being applied as a strategy:

> Statement 2: *I found it difficult to decide whether to skim (fast read) or read carefully the whole passage in order to answer the questions.*

The associated questionnaire statement related to the degree of difficulty that the students found in deciding which strategy would be most appropriate. A high number of test takers (64%) agreed that they found it difficult. Nevertheless, the statistical test indicated that a significantly higher number of those test takers who likely used discernment appropriately scored higher than those who did not apply this strategy (see Chapter 4 Section 4.5.3.2). This indicates that a significant number of test takers were

able to make prudent decisions regarding the appropriateness of careful or expeditious reading in a given context.

Nevertheless, there is some concern that there is an apparent contradiction between the results for the questionnaire statements of Discernment 1 and Discernment 2:

- Discernment 1 statement: *I found it difficult to decide whether to skim (fast read) or read carefully the whole passage in order to answer questions.*
- Discernment 2 statement: *I looked at the questions first before deciding whether to read the passage carefully or quickly.*

Two thirds of the students found it difficult to decide which strategy to use in *Discernment 1*, whereas in *Discernment 2* there was a high level of agreement regarding the strategy to be used which was that of looking at the question first before deciding which type of reading was appropriate. However, the questionnaire statement relating to *Discernment 1* seeks to elicit from the students whether or not they found it difficult. The fact of finding it difficult does not measure whether they correctly applied it or not. It is simply a finding, (not a surprising one), that the students found this strategy difficult to apply. However, Discernment 2 statement 26 clarifies the matter where the actual strategy is the focus of the question and here there is a high level of agreement (73%) among the test takers that they were appropriately applying the strategy.

It is admitted however that it is sometimes quite difficult to adequately incorporate discernment into a test of this nature. For example, it would be impractical to include ‘multiple texts’ on a theme or subject area where comprehending the contribution of each text would be required. This kind of situation calls for ‘an organising frame’ (Grabe, 2009) or ‘grouping relationships’ (Green, 2014), which the students would need at academic level but which is quite difficult to simulate in a test situation. This is actually one of the strategies which engaged readers use: “They read selectively according to goals” (Grabe, 2009, p.228) (see Chapter 2 Section 2.3).

It should be clear that more is involved here than target setting. Readers at academic level need to be able to decide how to effectively use the time available to them in reading. This involves a wise decision as to which passages or texts can be comprehended by expeditious reading and which texts require more careful reading and hence more time. This is not an automatic response but is, in fact, a highly developed metacognitive strategy (Purpura, 1999; Bachman and Palmer, 2010; see Chapter 2 Section 2.3), and one which students at academic level need to develop quickly during their first academic year. Green (2014) has emphasised the 'choice' the reader makes between expeditious and more careful reading as an important aspect of these metacognitive skills. The findings of this study indicate that the Level 4 students in Oman have been developing significant discernment skills which they can effectively draw on for the extensive reading which will be required at academic level.

5.3.3 Rubric 1

Rubric 1 concerned 4 test questions with an overall rubric:

PART D: Answer the questions. **(4 marks)**
10) What protects the beaches?
11) What did Zanzibar become the largest producer of?
12) Who moved his capital to Zanzibar?
13) What is the origin of most of the population in Zanzibar?

The rubric was somewhat vague about the level of detail required in the answer. For each of the 4 questions more than half of the test takers were agreed that they were unsure whether a short or longer response was required. This raises questions concerning the fairness of the test but

nevertheless a high number of test takers managed to answer correctly and the validity of the 4 questions is not in doubt. The reason for claiming this is that there is no ambiguity in the questions themselves and they turned out to be valid discriminators between those using the rubric as a guide and those who did not use the rubric appropriately. The only uncertainty surrounds the amount of detail that would be required and there might have been an issue of fairness in the sense that those who gave a longer answer but a correct one would still only merit one mark for their effort against those who gave a short correct answer but also received one mark. As mentioned in the analysis (Chapter 4 Section 4.5.3.2) some scripts were found which contained long detailed answers and this raises an issue of fairness. Apart from this set of questions, however, elsewhere rubrics are clear and unambiguous and provide guidance for the test takers. No evidence of ambiguity or unfairness was found apart from the one instance referred to (see Statements 19: *I found the test instructions easy to understand*, see Chapter 4 Section 4.5).

The questions concerned with rubric here are the same test questions which proved to be valid for measuring expeditious search (see Section 3.1 above). So looking at these questions from the perspective of the use of rubric, it appears to be a valid claim that the rubric guided the students and provided a fair framework so that no test taker was faced with ambiguity.

5.4 Summary

Based on students' test scores on the test paper used in this study in Oman, there was no evidence that choosing the intended cognitive process based on the cognitive processes under Component 1 (basic reading processes) resulted in a significantly higher correct score than choosing a different process (see Chapter 4 Section 4.5.3.1). In other words, getting a correct or incorrect answer had no relationship to the cognitive process used. This raises an issue of the validity of the test questions relating to the intended cognitive processes on Component 1. This means that the questions on the test paper that are purporting to measure certain cognitive processes were

not validly doing so. Hence, there is concern about the validity of inferences that may be made on the basis of scores for these particular questions. Additionally, it may not be appropriate to test students for skills which are yet insufficiently developed or practised. One such skill is expeditious reading, a construct for which there is sparse theoretical support beyond what is proposed in Khalifa and Weir (2009). The validity of expeditious reading as a construct may be questioned and consequently there is a need for caution in interpreting results based on variables purporting to measure it (See Section 5.2.1 above). The cognitive processes in Component 1 included *Scanning expeditiously*, *Careful global reading* and the knowledge and use of *Grammatical resources*. The evidence of this research is that Omani second language test takers were not effectively using these cognitive processes for comprehension.

However, on Component 2 (expeditious reading), evidence was found of an association between using the intended process and answering the test questions correctly (see Chapter 4 Section 4.5.3.2). Thus, on test items where expeditious reading processes were intended, those test takers who chose these processes tended to answer correctly significantly more often than those students who chose a different process. Accordingly, it is concluded that the inferences that are drawn from the scores relating to expeditious reading (Component 2) are valid. Thus, there is evidence that Omani second language test takers were effectively applying processes of expeditious reading such as *expeditious search reading*, the use of *discernment* and of *rubric* for comprehension.

The following chapter presents the analysis of academic texts and test tasks and is followed by a discussion chapter based on the findings. The final chapter forms the conclusion of the thesis where comparisons are drawn from the results of both sets of data analysis and discussions (see Chapter 3 Section 3.1). Convergences and divergences are noted and the implications for the cognitive processes which students used are inferred. These inferences lead to some recommendations being made as well as highlighting certain limitations of the study.

Chapter 6 Data Collection and analysis II: Automated analysis software and expert judges

6.1 Introduction

This chapter presents the results and findings of the data which was collected to answer the second and third research sub-questions (RSQ):

- How closely do the reading texts in LEE tests reflect those encountered at first year academic (FYA) level?
- How closely do the reading tasks in LEE tests reflect those encountered at first year academic (FYA) level?

Computational tools were utilized for analysing a selection of passages from Level Exit Exam (LEE) in the Foundation Program (FP) and text extracts from the First Year Academic texts (FYA) for comparative purposes. Additionally, where these computational tools were not directly applicable for measuring certain features, expert judges were employed to evaluate the texts and tasks (see Chapter 3 Section 3.5).

The computational tools included the following automated text analysis software: VocaProfiler, WordSmith and Coh-Metrix as described in the methodology chapter (see Chapter 3 Section 3.5). A total of 29 academic text extracts and 5 test texts were analysed. A larger selection of test texts would have been desirable but these were the only samples of previous tests which were available (see Chapter 3 Section 3.5.2).

Descriptive statistics were used to give an overview of the data and inferential statistics were used to investigate any significant differences in the data. Independent samples t-tests were conducted on 19 variables measuring Grammatical and Lexical resources (see Table 6.1 Code Book) and significant differences between academic texts and test passages were identified.

Table 6.1 Codebook

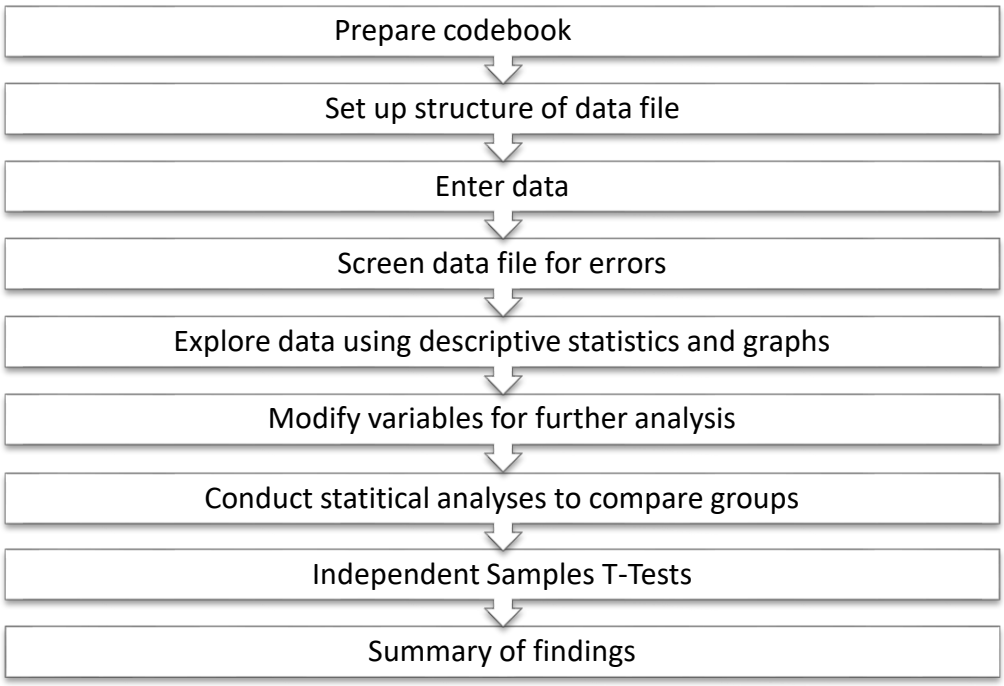
Variable #	Text analysis Software used	Variable Name	SPSS variable label	Coding instructions
		Text	Text type	Foundation Year Academic (FYA) = 1 Foundation Program year (LEE) = 2
1.	VocabProfile	V1	Average characters per word	
2.	WordSmith	V2	Standardised type-token ration	
3.	VocabProfile	V3	Lexical density	
4.	VocabProfile	V4	1000 word frequency	
5.	VocabProfile	V5	2000 word frequency	
6.	VocabProfile	V6	3000 word frequency	
7.	VocabProfile	V7	Frequency <15k	
8.	VocabProfile	V8	AWL level	
9.	Coh-Metrix	V9	Mean number of higher-level constituents per word	

10.	Coh-Metrix	G1	Average word per sentence
11.	Coh-Metrix	G2	Average sentences per paragraph
12.	Coh-Metrix	G3	Noun phrase incidence score per 1000 words
13.	Coh-Metrix	G4	Mean number of modifiers per noun phrase
14.	Coh-Metrix	G5	Mean Number of words before the main verb of main clause in sentences
15.	Coh-Metrix	G6	logical operators incidence score
16.	Coh-Metrix	R1	Flesch Reading Ease score (0-100)
17.	Coh-Metrix	R2	Flesch-Kincaid Grade level (0-16)
18.	Coh-Metrix	R3	Coh-Metrix L2 Readability
19.	Coh-Metrix	C1	Anaphor reference, all distances
20.	Coh-Metrix	C2	Argument overlap, all distances
21.	Coh-Metrix	C3	Proportion of content words that overlap between adjacent sentences

22.	Coh-Metrix	C4	LSA sentence to sentence adjacent mean
23.	Coh-Metrix	C5	LSA sentences all combinations mean
24.	Coh-Metrix	A1	Concreteness, mean for content words
25.	Coh-Metrix	A2	Concreteness, minimum in sentence for content words
26.	Coh-Metrix	A3	Mean hypernym values of nouns
27.	Coh-Metrix	A4	Mean hypernym values of verbs

For variables that were not directly amenable to scalar measurement, expert judges were employed to assess the texts drawing on their expertise. These variables were presented in a checklist which included a *Likert* scale for the judges to rate their judgments (see Chapter 3 Section 3.5.1). A total of 87 text extracts were selected from a range of subject specialisations and these were compared with the 15 test passages. The entire analytical process is outlined in the flowchart Figure 6.1:

Figure 6.1 Data analysis process



Adapted from Pallant (2010, p.28)

6.2 Descriptive statistics of automated analysis software

The variables were first coded to facilitate processing by the automated text analysis software (see Table 6.1). Variables V1, V3, V4, V5, V6, V7, and V8 were tested by Web VocabProfiler. Variable 2 was tested using WordSmith as this tool addresses the issue of varying text lengths which can affect the type-token ratio by using standardized text length of 250 words. The final variable of vocabulary (V9) was measured by using Coh-Metrix.

Grammatical and other lexical variables were analysed by Coh-Metrix. The full set of variables with a brief explanation of what each variable measures is presented in (see Table 6.1).

The academic texts were selected to represent the specializations and programs studied by the students at first year academic level. These descriptive statistics are presented in Table 6.2:

Table 6.2 Departments offering specialisation courses for First Year Academic

	Departments			
	Business Studies	**IT Department**	**Engineering Department**	**English Language Centre**
Courses	Principles of Marketing	Introduction to Networking	Physics 1	Technical Writing 1&2
	Principles of Microeconomic	Computer Hardware	Physics 2	
	Introduction to Business	Introduction to information system	Engineering work Mechanical	
	Managerial Statistics	Introduction to Web Technology	Engineering work Electrical	
	Principles of Accounting	Introduction to Operation Systems	Calculus	
	Job Search Techniques	Advanced IT	Computer Programming	
	Principles of Management	Applied Database	Chemistry	
	Business Communication	Introduction to Database	Engineering graphics	
		Applied Math		
		Pure Math		
		Managerial Statistics (IT)		
TOTAL	8	12	8	1

As can be seen from the Table 6.2, there are three main specializations: Business Studies with 8 individual courses, IT Department with 12 individual courses and Engineering with 8 individual courses. In addition, English is studied across all programs as one course divided into 2 sections, Technical Writing 1 & 2, each delivered in successive semesters.

6.3 Results and analysis for research sub-question 2:

How closely do the reading texts in LEE tests reflect those encountered at first year academic level?

The results from the various software analyses were loaded into SPSS and the two groups (First Year Academic (FYA) text extracts and Foundation Program test passages (LEE)) were compared using independent Samples t-tests (Table 6.3):

Table 6.3 Independent samples t-tests between FYA and LEE texts

	FYA		LEE		Levene's Test for Equality of Variance	Equal Variance Assumed/Not Assumed	t	df	Adj. Sig. (2-tailed)
	Mean	SD	Mean	SD	Sig.				
V1 Average characters per word	6.11	0.54	5.96	.09	.019	Not Assu.	1.41	31.949	.170
V2 Standardised type-token ration	41.29	1.83	32.98	10.65	.000	Not Assu.	1.740	4.041	.156
V3 Lexical density	0.60	0.10	0.60	0.04	.442	Assumed	-0.001	32	.999
V4 1000 word frequency	66.96	7.54	72.74	3.86	.090	Assumed	-1.662	32	.106
V5 2000 word frequency	12.44	3.50	10.94	1.72	.183	Assumed	929	32	.360
V6 3000 word frequency	8.37	3.28	5.05	1.04	.195	Assumed	2.222	32	.033
V7 Frequency <15k	0.30	0.19	0.56	0.64	.001	Not Assu.	-.884	4.125	.425
V8 AWL level	9.83	4.34	4.80	1.08	.055	Assumed	2.550	32	.016
V9 Mean number of higher-level constituents per word	2.21	0.34	1.75	0.31	.591	Assumed	2.788	32	.009
G1 Average word per sentence	11.23	2.35	15.54	1.89	.624	Assumed	-3.882	32	.000
G2 Average sentences per paragraph	1.56	0.37	6.86	1.93	.000	Not Assu.	-6.129	4.050	.003
G3 Noun phrase incidence score per 1000 words	403.50	41.14	388.41	24.17	.318	Assumed	.791	32	.435

G4 Mean number of modifiers per noun phrase	1.01	0.21	0.97	0.17	.544	Assumed	.393	32	.697
G5 Mean Number of words before the main verb of main clause in sentences	2.12	0.84	4.22	0.93	.548	Assumed	-5.118	32	.000
G6 logical operators incidence score	32.32	10.36	31.84	11.50	.991	Assumed	.093	32	.927
R1 Flesch Reading Ease score (0-100)	58.64	15.51	57.34	2.90	.014	Not Assu.	1.044	31.460	.304
R2 Flesch-Kincaid Grade level (0-16)	7.76	2.45	9.40	0.63	.025	Not Assu.	-3.069	26.427	.005
R3 Coh-Metrix L2 Readability	13.90	5.27	14.75	2.57	.061	Assumed	-.355	32	.725
C1 Anaphor reference, all distances	0.42	0.16	0.33	0.17	.868	Assumed	1.139	32	.263
C2 Proportion of content words that overlap between adjacent sentences	0.35	0.12	0.34	0.10	.337	Assumed	.124	32	.902
C3 Argument overlap, all distances	0.09	0.03	0.06	0.01	.076	Assumed	1.900	32	.066
C4 LSA sentence to sentence adjacent mean	0.30	0.09	0.21	0.04	.244	Assumed	2.172	32	.037
C5 LSA sentences all combinations mean	0.31	0.10	0.21	0.07	.271	Assumed	2.173	32	.037

A1 Concreteness, mean for content words	374.41	28.51	397.52	20.31	.468	Assumed	-1.729	32	.094
A2 Concreteness, minimum in sentence for content words	1.23	0.28	1.24	0.27	.773	Assumed	-.093	32	.927
A3 Mean hypernym values of nouns	6.38	0.52	5.47	0.79	.469	Assumed	3.359	32	.002
A4 Mean hypernym values of verbs	1.56	0.19	1.57	0.16	.558	Assumed	-.064	32	.955

As can be seen from the Adjusted Sig. (2-tailed), column 11 of the variables showed significant mean differences ($p < 0.05$) between the two groups. It was considered important to take into account whether equal variance was assumed or not assumed when interpreting the results of the t-tests due to the small sample size of test passages ($n = 5$) compared with the academic texts ($n = 29$). Levene's test for equality of variance significance was noted taking the 0.05 threshold whereby, for any significant result below that threshold, equal variance was not assumed and, above the 0.05 threshold, equal variance was assumed (Field, 2001; Foster, 2001; Pallant, 2010). However, it will be useful to present each variable in order, classified by the context validity features which they represented.

6.3.1 Grammatical resources: Vocabulary

Six of the nine vocabulary variables were shown to have no significant difference based on grouping by test passage or academic text. These variables were:

- V1 Average characters per word,
- V3 Lexical density,

- V4 First 1,000 word frequency,
- V5 Second 1,000 word frequency and
- V7 Frequency <15 k.
- V2 standardized type token ratio.

This indicates that, in the case of these six vocabulary indices, that the test passages (LEE) closely represented texts from first academic year (FYA). It is not surprising that, within the V4 *first* and V5 *second thousand word frequency*, no significant differences were found (A full discussion of the significance of most frequent word lists has been presented in Chapter 2 Section 2.7.3.6 Lexical Resources). However, in the third thousand most frequent words there was a significant difference (see V6 below). In the case of V1 *Average characters per word* and V3 *Lexical density*, LEE appears to match FYA. Finally, V7 *the words falling outside 15,000 word frequency,* were most likely to be technical words or proper nouns which would not be greatly different in LEE or FYA. In the case of technical words, these are often strictly defined within the specialisation context (see Chapter 2 Section 2.7.3.6).

The V2 *type-token ratio* results were obtained by means of WordSmith (Chapter 3 Section 3.5). The mean for the FYA texts was (M = 41.29) and for the LEE passages was (M = 32.98). A t-test on SPSS showed that the mean difference was not significant (t = 1.740, df = 4.041, p = .156, equal variance not assumed). Accordingly, the foundation test passages were found to be representative of first year academic texts in terms of type-token ratio.

Significant mean differences were found for three of the nine vocabulary measures. These were:

- V6 3,000 word frequency,
- V8 AWL level and
- V9 mean number of higher level constituents per word.

Although no significant differences were found for V4 *and* V5 *the first and second 1,000 most frequent words*, a significant difference was found on V6 *the third 1,000 most frequent words*. The mean scores for FYA texts (M = 8.37) and LEE passages (M = 5.05) were found to be significantly different (t

= 2.222; *df* = 32; *p* = 0.033). This result indicated that FYA texts scored significantly higher on vocabulary in the third most frequent 1,000 words than LEE. Thus, in relation to the 3,000 word frequency level, LEE passages were not representative of FYA texts.

For V8, *AWL scores*, indicating the percentage of sub-technical vocabulary, showed that FYA texts were significantly higher than LEE passages (*M* = 4.34 and 1.08, respectively). The independent samples t-test returned the following results: (*t* = 2.550; *df* = 32; *p* = 0.016). Thus, in relation to the percentage of sub-technical vocabulary, LEE passages were not representative of FYA texts.

The Coh-Metrix results for V9, *mean number of higher level constituents per word,* returned mean values of (*M* = 2.21 and *M* = 1.75 for FYA and LEE texts, respectively). These results were found to be significant on the independent samples t-test (*t* = 2.788; *df* = 32; *p* = 0.009). The test results showed that the FYA texts had a significantly higher mean score than the LEE passages. Thus, in terms of the mean number of higher level constituents per word, LEE passages were not representative of FYA texts. This is due to the fact that Foundation Program test passages have less specific terms than First Year Academic texts rendering it difficult for foundation texts to match academic texts on this variable. Foundation Programs tend to be general in nature whereas academic programs tend to be more specialised.

6.3.2 Grammatical Resources: Grammar

There were 6 variables (G1 to G6) representing grammar. On 3 of these, significant differences were found between FYA and LEE texts and no significant differences were found on the remaining 3 variables.
On the following 3 variables no significant differences were found between FYA and LEE texts:

- G3 Noun phrase incidence (Syntactic pattern density),
- G4 Mean number of modifiers per noun phrase (Syntactic complexity)
- G6 Logical operators' incidence score (Lexical diversity).

In the case of G3 (*noun phrase incidence)* FYA texts were found to be higher than LEE passages (*M* = 403.50 and *M* = 388.41, respectively) but the difference did not prove to be significant. On this evidence, LEE passages can be taken as reasonable representations of FYA texts in terms of noun phrase incidence.

The second non-significant variable was G4 *the mean number of modifiers per noun phrase*. This is an indicator of the relative complexity of noun phrases with academic texts being expected to score higher. However, the difference between FYA texts and LEE passages were found to be slight (*M* = 1.01 and *M* = 0.97, respectively) and this difference was not found to be significant. Thus, LEE passages were deemed to be representative of FYA texts in terms of the mean number of modifiers per noun phrase.

Similarly, G6 *logical operators incidence scores* (connectives) were found to be close for FYA and LEE texts (*M* = 32.32 and *M* = 31.84, respectively) and this small difference was not found to be significant. It can be concluded, therefore, that LEE passages are representative of FYA texts in terms of incidence of logical operators. In general, texts with a high density of logical operators are considered to be difficult for most readers (McNamara et al., 2005 in Green et al, 2010, p.198).

Significant differences were found between LEE and FYA texts on the following grammatical variables:

- G1 Average word per sentence
- G2 Average sentences per paragraph
- G5 Mean number of words before the main verb of the main clause (Syntactic complexity)

G1 *the average number of words per sentences* was found to be smaller for academic texts (*M* = 11.23) than for foundation test passages (*M* = 15.54). The difference was found to be significant ($t = -3.882$, $df = 32$, $p < 0.001$). Similar results were found for G2 *Average sentences for paragraphs* where academic texts (*M* = 1.56) were found to be less than foundation test passages (*M* = 6.86). This difference proved to be significant ($t = -6.129$, $df = 4.05$, $p = 0.003$). A full discussion of the significance of measures of average

word per sentence and average sentence per paragraph was presented in (Chapter 2 Section 2.7.3.5). These measures, however, are not the only factors to be considered when assessing the grammatical complexity of sentences of paragraphs as it is found that in academic texts shorter sentences and paragraphs are more challenging for the reader due to the opacity of language (see Chapter 2 Section 2.7.3.5). However, in terms of average number of words per sentence and also average number of sentences per paragraph, LEE passages were not representative of FYA texts.

Again, in relation to G5 *the mean number of words before the main verb of the main clause in sentences*, FYA texts (M = 2.12) were found to be less than LEE passages (M = 4.22). This difference was found to be significant (t = - 5.118, df = 32, p < 0.001). Generally, the remarks applying to opacity in the previous paragraph also operate for the number of words before the main verb. It would be expected that academic texts would have a higher number of words before the main verb and to have a high order of syntactic constituents. This is contrary to the results obtained here where the first year academic texts had significantly less words before the main verb than the foundation test passages and, consequently, LEE passages were not representative of FYA texts in this respect.
However, it should be borne in mind that the courses at academic level here were engineering, IT and business, all subjects which have high definition of vocabulary and less opaqueness than for academic subjects such as sociology, psychology, history…etc.

6.3.3 Readability

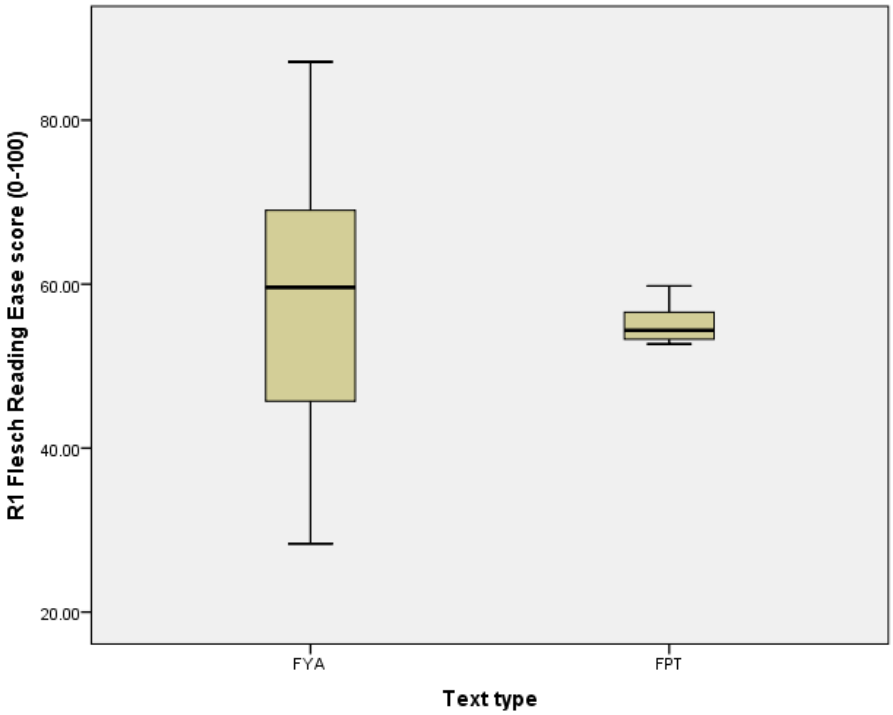
There were three measures of readability:

- R1 Flesch reading ease 0-100,
- R2 Flesch Kincaid Grade level 0-16 and
- R3 Coh-Metrix L2 readability.

Of these R1 and R3 were not found to differ significantly between foundation and first year academic texts. On R1 *Flesch reading ease 0-100* FYA texts

(*M* = 58.64) and LEE passages (*M* = 57.34) showed that both groups required some development of their skills as their scores were above 50. However, there is a notable difference in the standard deviations (*SD* = 15.51 and 2.90, respectively) showing a much wider range among the academic texts extracts as compared with the foundation test passages. This is shown in the boxplots in Figure 6.2:

Figure 6.2 Boxplots comparing FYA and LEE texts measured on the Flesch Reading Ease Scale

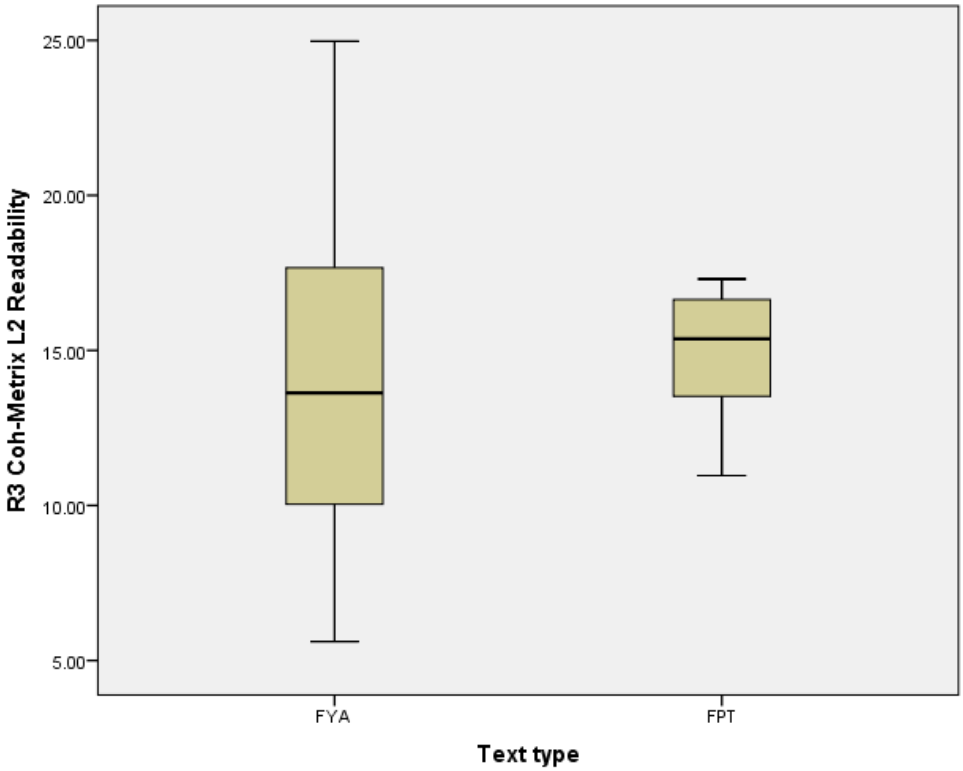


The boxplot for FYA texts shows that many of the texts are challenging (less than 50) whereas the LEE passages boxplot shows that none of the texts are challenging in this respect. Thus, using the Flesch reading ease scale, the LEE passages are not sufficiently challenging to match with many of the texts encountered at FYA, despite the fact that the mean differences were not found to be significant. The lower quartile of the FYA boxplot is located at approximately 45 on the scale indicating that 25% of these texts were much more challenging for academic students than those encountered at foundation level. In the LEE passages boxplot, the lower quartile and lowest measure were all above 50 which indicated that the test passages needed to be more challenging in terms of R1 *Flesch reading ease 0-100* if they were to

match the requirements of reading ease at First Year Academic level. However, it was noted that the upper quartile for FYA texts was approximately 70 indicating that 25% of texts at academic level were less challenging, although some authors have cautioned against over reliance on measures based on numbers of syllables, words and sentences for grammatical complexity in a text (See Chapter 2 Section 2.7.3.5).

Similarly R3, *the Coh-Metrix L2 readability* measure showed no significant difference between LEE and FYA texts. These results can be seen in boxplots in Figure 6.3:

Figure 6.3 Boxplots showing FYA and LEE scores for R3 Coh Metrix L2 Readability



A difference in the spread of the data for FYA and LEE was noted as being quite similar to that found for R1 in that LEE scores were much closer to the mean than FYA scores.

However, a significant difference was found on R2 *the Flesch Kincaid Grade level* between FYA and LEE texts with means (*M* = 7.76 and *M* = 9.40,

respectively). The t-test returned the following results (t = -3.069, df = 26.427, p = 0.005). Since scores 13 to 16 are considered to be equivalent to college level in the US school system, both of these scores fell below this threshold by as much as 3 or 4 levels. However, a question arises as to the applicability of the same threshold for texts being presented to L2 students in a non-English speaking context. Nevertheless, in the light of the t test results, it is concluded that LEE passages are not representative of FYA texts, even though both are several levels below that expected on the Flesch Kincaid Grade Level.

6.3.4 Cohesion

On the first 3 variables, C1, C2 and C3, representing various kinds of referential cohesion, the differences between LEE passages and FYA texts were not found to be significant. These were:

- C1 Anaphor reference
- C2 Argument overlap
- C3 Proportion of content words that overlap between adjacent sentences

Accordingly, these 3 types of referential cohesion occurring in test passages were found to be representative of similar types occurring in first year academic texts.

On the remaining 2 variables of referential cohesion, (C4 and C5), significant differences were found between LEE passages and FYA texts.

C4 *Latent semantic analysis (LSA)* is a measure of semantic or conceptual similarity between text excerpts. Such similarities are understood as aiding comprehension. C4 measures sentence to sentence cohesion for adjacent sentences. The measure for FYA texts was (M = 0.30) and for LEE passages (M = 0.21). The means were compared using the independent samples t-test and the following results were obtained: (t = 2.172, df = 32, p = .037). In terms of latent semantic analysis, first year academic texts were found to be significantly higher than foundation test passages. Thus, LEE passages were

not representative of FYA texts in terms of LSA. C5 is similar to C4 except that the comparison is for latent semantic analysis across all sentences in a passage. For FYA texts, the mean score was (*M* = 0.31) and for LEE passages (*M* = 0.21). The t-test results were as follows: (*t* = 2.173, *df* = 33, *p* = 0.037). The t test result was similar to that obtained for C4. FYA texts were found to be significantly higher in LSA than LEE passages. Thus, LEE passages are not representative of FYA texts in terms of latent semantic analysis across sentences.

6.3.5 Abstractness

Four variables were used to measure levels of abstraction. On three of these no significant differences were found between FYA texts and LEE passages:

- A1 Concreteness for content words, mean,
- A2 Concreteness minimum in sentence for content words, mean and
- A4 Mean hypernym value of verbs.

All three measures are different ways of measuring how abstract a word is. However, in relation to these three measures - A1, A2 and A4 - LEE passages were found to be representative of FYA texts.

On one variable a significant difference was found between LEE passages and FYA texts. This variable was:

- A3 Mean hypernym value of nouns.

In contrast to A4 (*Mean hypernym value of verbs)*, the mean difference for A3 - *Mean hypernym value of nouns* - for FYA texts (*M* = 6.38) and for LEE passages (*M* = 5.47) was found to be significant (*t* = 3.399, *df* = 32, *p* = 0.002). This significant difference was found in the direction of FYA texts having a significantly higher mean hypernym value for nouns than was found for LEE passages. Thus, it is concluded that first year academic texts tended to have a higher hypernym value for nouns compared to foundation test passages and, consequently, LEE passages were not representative of FYA texts in terms of hypernym values for nouns.

6.4 Descriptive statistics for expert judges analysis

Some of the variables were not directly amenable to scalar measurement and these were assessed qualitatively by expert judges. The data relating to the ratings of these expert judges were coded and entered into SPSS. These are presented in Table 6.4:

Table 6.4 Code book for features assessed by expert judges

Variable #	SPSS variable label	Coding instructions
	Text type	Foundation Year Academic (FYA) = 1 Foundation program year (LEE) = 2
1.	overall text purpose	Referential = 1 Conative = 2 Emotive = 3 Poetic = 4 Phatic = 5
2.	Writer-reader relationship	Audience addressed = 1 Audience invoked = 2
3.	Discourse Mode Genre	Textbook =1 Magazine/newspaper article = 2 Research or academic journal article = 3 Report = 4
4.	Rhetorical task	Exposition =1 Argumentation/persuasion/evaluation=2 Historical/biographical/autobiographical narrative=3
5.	Pattern of exposition	Define=1 Describe=2 Elaborate=3 Illustrate=4 Compare or contrast=5 Classify=6 Cause or effect=7 Problem/solution=8 Justify=9
6.	Rhetorical organisation	Explicit=1 Somewhat explicit=2 Neutral=3 Somewhat not explicit=4 Not explicit=5
7.	Functional resources	Ideational=1 Manipulative=2 Heuristic=3 Imaginative=4
8.	Grammatical resources Grammar	Mainly simply sentences=1 Balance of simple and compound sentences=2 Mostly compound sentences=3 A balance of compound and complex sentences=4 Mostly complex sentences=5
9.	Grammatical resources Cohesion	Explicit=1 Somewhat explicit=2 Neutral=3 Somewhat not explicit=4 Not explicit=5
10.	Content knowledge	General=1 Somewhat general=2

		Neutral=3 Somewhat specific=4 Specific=5
11.	Cultural background	Cultural neutral=1 Somewhat cultural neutral=2 Neutral=3 Somewhat cultural specific=4 Cultural specific=5
12.	Language background	First language background neutral=1 Somewhat first language background neutral=2 Neutral=3 Somewhat first language background specific=4 First language background specific=5
13.	Religion knowledge	Religion neutral=1 Somewhat religion neutral=2 Neutral=3 Somewhat religion specific=4 Religion specific=5
14.	Channel of Presentation	Appropriate=1 Somewhat appropriate=2 Neutral=3 Somewhat not appropriate=4 Not appropriate=5
15.	Text length	Appropriate=1 Somewhat appropriate=2 Neutral=3 Somewhat not appropriate=4 Not appropriate=5

A text was selected for each of the 29 courses (see Table 6.2 above) and each text extract was assessed by three different judges. Additionally, selections from 5 foundation test passages were chosen and each passage was assessed by three judges. Thus, there was a total of 102 observations and judgments for each of the 15 contextual features. This data is presented in Table 6.5:

Table 6.5 Number of observations for each contextual feature

Contextual features variables	N
overall text purpose	102
Writer-reader relationship	102
Discourse mode Genre	102
Rhetorical task	101
Pattern of exposition	102
Rhetorical organization	99
Functional resources	101
Grammatical resources Grammar	102
Grammatical resources Cohesion	102
Content knowledge	102
Cultural background	102
Language background	102
Religion knowledge	102
Channel of presentation	95
Text length	95

For most of the variables there were 102 different assessments. In a few cases there were missing values where a judge did not assign an option but left the particular question blank. However, there were not many missing values, the highest being 7 for both channel of presentation and text length.

The rate of agreement between the judges on the various criteria was calculated and the results are shown in Table 6.6:

Table 6.6 Rates of agreement between the judges for each contextual feature

	Exact	+/-1
overall text purpose	90%	100%
Writer-reader relationship	90%	100%
Discourse mode	56%	79%
Functional resources	66%	76%
Grammatical resources	51%	89%
Content knowledge	72%	85%
Channel of presentation	46%	73%
Text length	55%	91%

There were 8 criteria represented by 15 different features, as Discourse mode was represented by 4 variables, Grammatical resources by 2 variables and Content knowledge by 4 variables.

In general, Table 6.6 shows an acceptably high rate of agreement between the judges (>50%) with the exception of Channel of presentation (46%). However, in the latter case, when the rate of agreement is extended by +/- 1, this rises to 73% and since the mode was 1 *appropriate*, category 2 accounted for an additional 27% of agreement. In 73% of the evaluations, the judges agreed that the Channel of presentation was either *appropriate* or *somewhat appropriate*.

Having established that there was an acceptable degree of agreement between the judges, the non-parametric Mann-Whitney U test was run to discover whether there were any significant differences between the scores

for the academic texts and the test passages. The results of the Mann-Whitney U test are presented in Table 6.7:

Table 6.7 Mann Whitney U Test results for expert judges' scores on each variable for LEE and FYA scores

	overall text purpose	Writer-reader relationship	Discourse Mode Genre	Rhetorical task	Pattern of exposition	Rhetorical organisation	Functional resources
Mann-Whitney U	627.500	628.500	439.000	564.500	595.500	396.500	394.500
Wilcoxon W	747.500	748.500	4267.000	4305.500	4423.500	3966.500	514.500
Z	-.458	-.440	-2.322	-.997	-.559	-2.756	-2.872
Asymp. Sig. (2-tailed)	.647	.660	.020	.319	.576	.006	.004

	Grammatical resources Grammar	Grammatical resources Cohesion	Content knowledge	Cultural background	Language background	Religion knowledge	Channel of Presentation	Text length
Mann-Whitney U	414.000	541.000	245.500	633.000	529.000	592.500	387.000	524.000
Wilcoxon W	4242.000	4369.000	365.500	753.000	649.000	712.500	3708.000	3845.000
Z	-2.496	-1.154	-4.039	-.278	-1.340	-1.216	-2.019	-.508
Asymp. Sig. (2-tailed)	.013	.248	.000	.781	.180	.224	.043	.611

Significant differences were found on the following variables:

- Discourse mode – Genre (p = .020)
- Discourse mode – Rhetorical organization (p = .006)
- Functional resource (p = .004)
- Grammatical resource – Grammar (p = .013)
- Content knowledge – Subject specificity (p <.001)
- Channel of presentation (p = .043)

However, each variable in turn will now be commented on and the findings of the Mann-Whitney U test will be indicated. The significant findings will be presented first. In order to determine the direction of the significant results, Pallant (2010) was followed by first referring to the table of ranks looking at the mean rank column for an indication of which group (in this study academic texts and test passages) was higher and then reporting the median in both cases. The counts for each variable by text type will also be included.

6.4.1 Discourse mode – Genre (p=.020)

Identify the most appropriate category for the text.				
Options	**1** text book	**2** magazine/newspaper article	**3** research/academic journal article	**4** report

Table 6.8 shows the decisions of the judges on the *Discourse mode - Genre* of the academic texts and the foundation test passages:

Table 6.8 Judges' ratings for genre by text type

		text type		Total
		Academic text extract	**Foundation test passage**	
Discourse mode Genre	textbook	59	4	63
	Magazine/newspaper article	6	6	12
	research or academic journal article	16	4	20
	report	6	1	7
Total		87	15	102

The table shows that the majority of the academic text extracts were judged to be *textbooks* (*n* = 59) or *research or academic* (*n* = 16) types. In contrast, judgments were more varied for foundation test passages, the majority being judged as *magazines* or *newspapers articles* types.

In the Mann-Whitey U test, FYA text ratings were found to be significantly lower than those for LEE passages (Mean Rank 49.05 and 65.73, respectively) with a median (*Md* = 1, *textbook*) for academic texts and median (*Md* = 2, *magazine or newspaper articles*) for test passages. Thus, while FYA texts were deemed to be *textbook type*, LEE passages were more frequently judged to be *magazine or newspaper articles* type and, accordingly, were not reflective of the text types in academic texts in terms of discourse mode genre.

6.4.2 Rhetorical organization (*p* = .006)

The organisational structure of the text is…					
	1	**2**	**3**	**4**	**5**
Options	Explicit				Not explicit

The judges' decisions regarding the explicitness of the organisational structure of the texts are presented in Table 6.9:

Table 6.9 Judges' ratings for Discourse mode - Rhetorical organisation by text type

		text type		Total
		Academic text extract	Foundation test passage	
Rhetorical organisation	explicit	61	5	66
	somewhat explicit	18	9	27
	neutral	2	0	2
	somewhat not explicit	0	1	1
	not explicit	3	0	3
Total		84	15	99

Generally, the judges found that rhetorical organization was explicit in both cases, the difference being in the degree of explicitness. Academic texts were mostly found to be *explicit* (*n* = 61) or *somewhat explicit* (*n* = 18) whereas foundation test passages were only judged to be *explicit* in 5 cases and somewhat *explicit* in 9 cases.

In the Mann-Whitney U test, FYA texts ratings were found to be significantly lower than LEE passages (Mean rank = 47.22 and 65.5, respectively). The median for academic texts was (*Md* = 1, *explicit*) and median for foundation test passages was (*Md* = 2, *somewhat explicit*). Thus, the findings indicated that in terms of rhetorical organisation, LEE passages were not representative of FYA texts.

6.4.3 Functional resources (*p*=.004)

Identify the most appropriate category for the text.				
	1	**2**	**3**	**4**
Options	Ideational	Manipulative	Heuristic	Imaginative

The judges' assessments in relation to the appropriate category for each text by Functional resources are presented in Table 6.10:

Table 6.10 Judges' ratings for Functional resources by text type

		text type		Total
		Academic text extract	**Foundation test passage**	
Functional resources	ideational	15	9	24
	manipulative	9	0	9
	heuristic	61	6	67
	imaginative	1	0	1
Total		86	15	101

The majority of FYA texts were classified as *heuristic* (language for extension of knowledge through various devices such as problem-solving, *n* = 61) followed by *ideational* (based on real-life experiential knowledge, *n* = 15). In

contrast, the majority of LEE passages were classified as *ideational* (*n* = 9) followed by *heuristic* (*n* = 6). The result of the Mann-Whitney U test showed that FYA texts mean rank score was 53.91, significantly higher than LEE passages 34.30. Academic text extracts were more frequently judged to be *heuristic* (*Md* = 3) in contrast with foundation test passages which were more frequently classified as *ideational* (*Md* = 1). Thus, in terms of functional resources, LEE passages were found to be not representative of the heuristic type of text, which was more characteristic of FYA texts.

6.4.4 Grammatical resources: Grammar (*p* = .013)

The sentences in the text are…					
Options	**1** mainly simple sentences	**2** a balance of simple and compound sentences	**3** mostly compound sentences	**4** a balance of compound and complex sentences	**5** mostly complex sentences

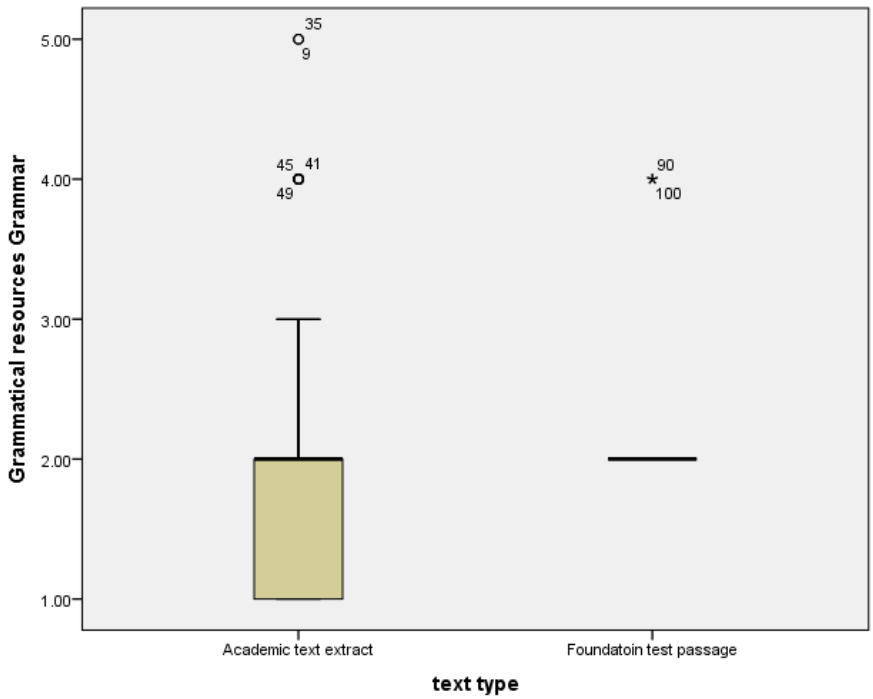
The results of the judges' assessments of *Grammatical resources - Grammar* represented in academic and foundation texts were classified on a continuum ranging from simple sentences to complex sentences as shown in Table 6.11:

Table 6.11 Judges' ratings for Grammatical resources (Grammar) by text type

		text type		Total
		Academic text extract	**Foundation test passage**	
Grammatical resources Grammar	mainly simple sentences	36	0	36
	balance of simple and compound sentences	40	13	53
	mostly compound sentences	2	0	2
	a balance of compound and complex sentences	7	2	9
	mostly complex sentences	2	0	2
Total		87	15	102

FYA texts were most frequently classified as a *balance of simple and compound sentences* (*n* = 40) or *mainly simple sentences* (*n* = 36). In contrast, LEE passages were most frequently found to have *a balance of compound and complex sentences* (*n* = 13). The Mann-Whitney U test revealed that academic texts were significantly lower (Mean Rank = 48.76) than for foundation test passages (Mean Rank = 67.40) on Grammatical resources (Grammar). However, the median score in either case was (*Md* = 2). A rating of 2 on the judges' checklist indicated "*a balance of simple and compound sentences*" as the most appropriate descriptor of the grammatical resources characterizing the academic texts and test passages. Boxplots graphical output was obtained and these are presented in Figure 6.4:

Figure 6.4 Boxplots on Grammar for academic text extracts and foundation test passages



For FYA texts both the median and upper quartile lies at 2 whereas practically all the data for LEE passages are located at 2. However, there is more variation in FYA texts with 25% of the data higher than 2 and 25% lower than 2, with three outliers at 4 and two outliers at 5. In the case of LEE passages all the data is at 2 except for two outliers at 4. The interpretation of the findings is that there is a significant difference in the way in which grammatical resources are used to convey meaning in FYA texts and LEE passages. This difference is in terms of a wider variation in the judges' assessments of grammatical resources used in academic texts than that of foundation tests. This difference is in both directions with some academic texts being assessed as containing mainly simple sentences and others being assessed as consisting mostly of compound sentences and, in the case of the outliers, more complex sentences. By contrast, there is scarcely any variation in the foundation test passages which were assessed as mainly consisting of a balance of simple and compound sentences and therefore LEE passages were not representative of the variety of sentence types found in FYA texts.

6.4.5 Content knowledge ($p < .001$)

Is the topic of the text of general interest or does it require subject specific knowledge on the part of the reader?					
Options	**1** (general	**2**	**3**	**4**	**5** (specific)

The results of the judges' decisions regarding the content knowledge nature of the text are presented in Table 6.12 on which the judges used the continuum from (1 = *general*) to (5 = *specific*) to rate each text.

Table 6.12 Judges' ratings for Content knowledge by text type

		text type		Total
		Academic text extract	**Foundation test passage**	
Content knowledge	general	13	7	20
	somewhat general	4	1	5
	neutral	15	6	21
	somewhat specific	12	1	13
	specific	43	0	43
Total		87	15	102

Table 6.12 shows that FYA texts were most frequently rated as *specific* (*n* = 43) whereas LEE passages tended to be most frequently rated as *general* (*n* = 7). However, in a high number of cases both FYA texts (*n* =15) and LEE passages (*n*= 6) were rated as neutral. The Mann-whiney U test revealed that FYA texts had a significantly higher mean rank score (56.18) than LEE passages (24.37). The median score for FYA texts was (*Md* = 4, *somewhat specific*) and for LEE passages was (*Md*= 2, *somewhat general*). Thus, over 50% of the judges' decisions on academic texts categorized them as either

specific or *somewhat specific* whereas their decisions for foundation test passages categorized over 50% as either *general* or *somewhat general*. In other words, academic texts generally were found to require greater subject specific knowledge than foundation test passages. Consequently, LEE passages did not reflect the demand for subject specific knowledge on the part of the reader which would be required for comprehension of FYA texts.

6.4.6 The remaining variables

On the remaining variables, no significant differences were found between FYA texts and LEE passages. *Overall text purpose* and *writer-reader relationship* had very close mean rank scores and both had equal medians (*Md* = 1). Thus, the judges agreed that both academic texts (78 out of 87 texts) and foundation test passages (14 out of 15 test passages) tended to be *referential* (intended to inform) in terms of overall text purpose demonstrating that LEE passages reflected the overall text purpose to be found in FYA texts. Similarly, the judges identified target readers as *audience addressed* (the intended reader) in both groups (*n* = 78 out of 87 for FYA texts and *n* = 14 out of 15 for LEE passages), again demonstrating that in terms of targeted audience, LEE passages were representative of FYA texts.

For Discourse mode (Rhetorical task), the judges identified *exposition* as the most appropriate category for both FYA and LEE texts (*n* = 64 out of 86 texts and *n* = 10 out of 15 passages, respectively). Thus, LEE passages were reflective of the rhetorical task characteristic of FYA texts.

However, for Discourse mode (Pattern of exposition) there was a difference in that the judges identified *describe* as the pattern used in the FYA texts (*n* = 32 texts out of 87) whereas they were divided between *describe* and *elaborate* as the pattern used in LEE passages (*n* = 6 out 15 for each). This difference, though, was not found to be significant in the Mann-Whitney U test. Because of this, LEE passages were therefore found to be reasonably representative of FYA texts although there was a greater range of pattern of exposition required at academic level as shown in Table 6.13:

Table 6.13 Judges' ratings for Discourse mode - Pattern of exposition by text type

		text type		Total
		Academic text extract	Foundation test passage	
	define	13	0	13
	describe	32	6	38
	elaborate	13	6	19
	illustrate	21	1	22
Pattern of exposition	compare of contrast	0	2	2
	classify	3	0	3
	cause or effect	3	0	3
	problem/solution	1	0	1
	justify	1	0	1
Total		87	15	102

There was also a median difference found for *Grammatical resources Cohesion*, with FYA texts having a median rating (*Md* =1) and LEE passages having a median rating (*Md* = 2). This difference was in terms of explicitness with First Year Academic texts being judged to be *more explicit* (*n* = 44 out of 87 texts) than Foundation Program test passages which were only rated as *somewhat explicit* (*n* = 7 out of 15 passages). However, the difference was found to be non-significant in the Mann-Whitney U test. It is therefore concluded that, in terms of explicitness of grammatical resources (cohesion), LEE passages were found to be representative of FYA texts.

The judges' assessments of *Content knowledge* for (*Cultural background, Language background and Religion knowledge*) showed very similar ratings on FYA and LEE texts with a median (*Md* = 1) on each of these three variables. A median score (*Md* = 1) signified cultural neutrality in the case of

Cultural background, first language background neutrality in the case of Language background and religion neutrality in the case of Religion knowledge. Thus, on these three variables, no significant differences were found between the judges' ratings in each case. It is therefore concluded that, in terms of Cultural background, Language background and Religion background, LEE passages were representative of the neutrality found in FYA texts.

6.5 Results and findings for research sub-question 3:

How closely do the reading tasks in LEE tests reflect those encountered at first year academic level?

In this study, task setting was investigated by examining two features, namely channel of presentation and text length. These features were not directly amenable to software analysis; instead, the judgments of experts were called upon to match these features as they were found in test tasks with comparators drawn from academic texts. It was not possible to examine other features of task setting using the method adopted in this study. However, the findings for task setting, based on these two parameters, are presented in the following sections.

6.5.1 Channel of presentation (*p*=.043)

The channel for the target situation requirements of the students being tested is…					
Options	**1** (appropriate)	**2**	**3**	**4**	**5** (not appropriate)

The judges' assessments of the appropriateness of the channel of presentation for the texts are presented in Table 6.14 where a continuum scale from (1 = *appropriate* to 5 = *not appropriate*) was used.

Table 6.14 Judges' ratings for Channel of presentation by text type

		text type		Total
		Academic text extract	**Foundation test passage**	
Channel of Presentation	Appropriate	41	3	44
	somewhat appropriate	21	5	26
	neutral	11	4	15
	somewhat not appropriate	4	0	4
	not appropriate	4	2	6
Total		81	14	95

This was one of the task setting features where it was possible to apply judges' assessments by comparing academic texts and test passages. A general view of the data (Table 6.14) shows that, in both cases, the judges found the Channel of presentation to be *appropriate* (FYA texts *n* = 62 and LEE passages *n* = 8). However, using a 5 point *Likert* scale, there was a subtle but significant difference.

The Mann-Whitney U test revealed that FYA tasks were found to have a significantly lower rating (a mean rank score of 45.78) than LEE test tasks (60.86). FYA tasks were most frequently rated as *appropriate* (*Md* = 1) whereas LEE tasks were most frequently rated as *somewhat appropriate* (*Md* = 2). The difference therefore was in terms of the degree of appropriateness of the channel of presentation as an aid to comprehension. The judges' assessments showed that channel of presentation was more appropriate as an aid to comprehension in academic texts than in test tasks where it was judged to be somewhat appropriate. Therefore, LEE test tasks, although somewhat appropriate in terms of channel of presentation, did not fully reflect the requirements of channel of presentation found in FYA tasks.

6.5.2 Text length

The text length for the target situation requirements of the students being tested is…					
Options	**1** (appropriate)	**2**	**3**	**4**	**5** (not appropriate)

Finally, on the task setting feature of text length, a median difference was found regarding the judgment as to whether the passages were of sufficient length and size for the purposes of comprehension. Academic text extracts were found to be *appropriate* (*Md* = 1) whereas foundation test passages were only adjudged to be *somewhat appropriate* (*Md* = 2) although this difference was found to be non- significant on the Mann-Whitney U test. Thus, it is concluded that, for purposes of comprehension, that LEE text lengths were generally representative of FYA passages.

6.6 Summary

The purpose of this chapter was to address the two research sub-questions by various methods in order to assess how representative foundation test texts and test tasks were of the texts encountered at first year academic level. Samples of texts from foundation tests and from academic texts across 29 subject area text books were selected and subjected to a number of text analyses using appropriate analytical tools. A number of features, which were not directly amenable to scalar measurement, were assessed by expert judges.

In terms of vocabulary measures, foundation test passages were found to be representative of first year academic texts for the following variables:

- V1 *Average number of characters per word*
- V2 *Standardised type-token ratio*
- V3 *Lexical density*
- V4 *Percentage of words occurring in the most frequent 1,000 words*

- V5 *Percentage of words occurring in the second most frequent 1,000 words*
- V7 *Percentage of words falling outside the 15,000 word frequency level*

Significant differences were found for the remaining three vocabulary measures. Foundation test passages were found to be unrepresentative of words in the third (V6) *1,000 most frequent word list*; for (V8) *AWL scores*, for percentage of words from sub-technical vocabulary, and for (V9) *average number of higher level constituents per word*. On each of these three measures, LEE passages were found to be unrepresentative of FYA texts.

Foundation test passages were found to be representative of First Year Academic texts for the following grammatical measures:

- G3 *Proportion of words included in noun phrases*
- G4 *Number of modifiers per noun*
- G6 *Logical operators incidence*

However, scores were found to be significantly higher for Foundation Programme test passages than for first year academic texts on the following grammatical measures:

- G1 *Average word per sentence*
- G2 *Average sentence per paragraph*
- G6 *Lexical density*

Thus, sentences and paragraphs tended to be longer in LEE passages and were found to be more difficult in terms of density of logical operators than was the case with the academic texts. On measures of readability, LEE passages were representative of FYA texts on the R1 *Flesch Reading Ease measure* and on R3 *the Coh-Metrix L2 Readability* scale. However, LEE passages were found to be rated significantly lower on R2 *the Flesch-Kincaid Grade Level* than was the case for FYA texts.

Foundation Program tests passages were found to be representative of academic texts for three of the referential cohesive devices, namely C1 *anaphor reference*, C2 *argument overlap* and C3 *proportion of content that overlap between adjacent sentences*. However, LEE passages were found to

be unrepresentative of FYA texts in terms of latent semantic analysis either for C4 *adjacent sentences* or for C5 *a passage as a whole*.

On word information, related to measures of concreteness or abstraction in words, LEE passages were found to be representative of FYA texts in terms of A1 *mean* and A2 *minimum concreteness of word content* and also for *mean hypernym value of main verbs*. A significant difference was found, however, for *mean hypernym value of nouns* where LEE passages were found to have significantly lower hypernym values than FYA texts and were not, thus, sufficiently representative.

For assessments of linguistic demands and task setting, LEE passages were found to be representative of FYA texts for the following features:

- Overall Text Purpose
- Writer-reader relationship
- Pattern of Exposition
- Grammatical Resource (Cohesion)
- Content Knowledge (Cultural, language background and Religion)
- Text length.

Significant differences were found for *Genre* where LEE passages tended to be representative of magazine and newspaper excerpts rather than of FYA texts. LEE passages were also found to be *rhetorically less explicit* than FYA texts. In terms of *functional resources*, LEE passages tended to be more ideational and less heuristic than was the case for FYA texts. On *grammatical resources (grammar),* LEE passages tended to represent a mixture of simple and compound sentences but lacked the variation in complexity which was characteristic of FYA texts. With regard to subject specificity in content knowledge, LEE passages tended to rely on more general content knowledge rather than subject specificity which was more characteristic of FYA texts. Finally, the task setting of channel of presentation was found to be less appropriate for LEE than for FYA texts. These findings are critically discussed with reference to the relevant literature in Chapter 7.

Chapter 7 Discussion II: Automated analysis software and expert judges

7.1 Introduction

This chapter discusses the findings from the analysis of the data generated by the automated software and expert judges' assessments (Chapter 6). The findings are critically examined in the light of the relevant literature.

The aims and objectives of Chapter 6 were to address the second and third research sub-questions (RSQ):

- How closely do the reading texts in LEE tests reflect those encountered at first year academic level?
- How closely do the reading tasks in LEE tests reflect those encountered at first year academic level?

The chapter proceeds by following the order of the features as they have been presented in the findings in Chapter 6.

7.2 Grammatical resources

Grammatical resources were examined quantitatively, based on measurements of four contextual features namely, *vocabulary*, *grammar*, *readability* and *cohesion*, and qualitatively, based on *cohesion* and *grammar* (see Chapter 2 Section 2.7.3.5, Chapter 6 Section 6.3.1 and Section 6.3.2,). The findings related to each of these are now discussed in the following subsections.

7.2.1 Grammatical resources: Vocabulary

9 variables were used to evaluate vocabulary for the purposes of comparison between LEE passages and FYA texts. 6 of these were found to be non-significant and 3 were significant as shown in Table 7.1:

Table 7.1 Grammatical resources: Vocabulary

Variables with NO significant difference	Variables with significant difference
V1 *Average characters per word*	V6 *3000 word frequency*
V3 *Lexical density*	V8 *AWL level*
V4 *First 1000 word frequency*	V9 *Mean number of higher level constituents per word*
V5 *Second 1000 word frequency*	
V7 *Frequency <15 k*	
V2 *Standardized type token ratio*	

The six variables which were found to reveal non-significant differences between the LEE passages and FYA texts indicated that, on these measures, the tests were broadly representative of the academic texts in the first year. However, it must be borne in mind that some of these variables, for example V1 *Average characters per word* and V3 *Lexical density* are often crude measures to be used as rules of thumb, as suggested in (Chapter 2 Section 2.7.3.6), which do not take into account the context in which the word occurs.

V2 *Standardised type-token ratio* is a measure of how demanding a passage is in terms of the number of different words required for comprehension. There was no significant difference on this measure between LEE passages and FYA texts.

Prima facie, it appears somewhat anomalous that on V7 *Frequency <15,000* no significant difference was found between LEE passages and FYA texts whereas on V6, *the third most 1,000 frequent* words a significant difference was found. V7 is measuring the mean number of words lying outside the 15,000 word threshold and probably consists of technical terms or names of cities, persons or other proper nouns. The LEE passages did contain a list of places and names which would count as proper nouns and therefore showed a higher percentage (0.56%) than FYA texts (0.3%). Nevertheless, this

difference proved to be non-significant (see Table 6.4 in Chapter 6). V6, on the other hand, measures word frequency according to a prescribed list of the third most frequent 1,000 words. V6 is closely linked with V8 *AWL level* (sub-technical) and V9 *Mean number of higher level constituents per word*, the other two significant variables. However, this result is explained by the fact that many academic texts include very specific terms not usually characteristic of non-academic texts. In other words, the genre type of these texts reflects the vocabulary characteristic of academic texts whereas LEE passages use a more general vocabulary typical of magazines/newspapers articles. This result is consistent with the findings for discourse mode as genre in Section 7.3 below where a significant difference is also reported.

In relation to the discussion of V6, Koda (2004) has argued in favour of L2 readers learning the most frequent 2,000 words claiming that these constituted approximately 80% of core vocabulary. The implications are that the next 1,000 most frequent words would be quickly acquired by inferring the meaning from the context. This would appear to be borne out by the findings of this research as there was no significant difference between LEE passages and FYA texts in terms of the first and second 1,000 most frequent words while a significant difference was found on the third 1,000 most frequent words and V8 (sub-technical words) as these were more represented in academic texts than in LEE passages. However, Khalifa and Weir (2009) have argued that, while word lists are useful for vocabulary building at lower levels, they may be less useful at higher levels. Instead, they recommend raising the threshold to 5,000 words on the grounds that the 2,000 words threshold was based on a 95% vocabulary recognition level whereas 98% coverage was argued to be more appropriate for L2 readers advancing to academic level. The findings of this research support this argument as the results of V6, V8 and V9 showed that FYA texts were significantly higher than LEE passages for L2 readers in a non-English speaking context.

In summary, in terms of vocabulary, the LEE passages were found to be representative of academic texts, except on those variables that represented a line of demarcation between L2 readers at foundation level and L2 readers

at first year academic level. Accordingly, on those variables, there was an under-representation in LEE. It was due to such under-representation in terms of technical vocabulary (V6, V8 and V9) that the test content lacked validity which would raise doubt over decisions about students' performances based on inferences drawn on V6, V8 and V9. Minimising construct under-representation and construct irrelevant variance was discussed in the literature as necessary for the avoidance of invalid score interpretation (see Chapter 2 Section 2.6).

7.2.2 Grammatical resources: Grammar

The second subdivision of Grammatical resources is *grammar* which is represented by 6 variables denoted by G1 to G6 (see Chapter 6 Section 6.3.2). Three of these variables showed significant differences between LEE passages and FYA texts and differences in the remaining 3 were found to be non-significant. These are set out in Table 7.2:

Table 7.2 Grammar variables assessed by automated software analysis

Variables with NO significant difference	Variables with significant difference
G3 *Noun phrase incidence (Syntactic pattern density)*	G1 *Average words per sentence*
G4 *Mean number of modifiers per noun phrase (Syntactic complexity)*	G2 *Average sentences per paragraph*
G6 *Logical operators' incidence score (Lexical diversity)*	G5 *Mean number of words before the main verb of the main clause (Syntactic complexity)*

In terms of G3 *Noun phrase incidence (Syntactic pattern density),* LEE passages were found to have a lower mean density than FYA texts but this difference was not significant. For G4 *Mean number of modifiers per noun phrase (Syntactic complexity)* and G6 *Logical operators' incidence score (Lexical diversity)* there were only slight differences between LEE passages

and FYA texts and these differences were not significant. Accordingly, in terms of these three measures, LEE passages were found to be representative of FYA texts. This was a reassuring and salutary finding due to the importance of syntactical knowledge for comprehension. In fact, the importance of syntactical knowledge has been emphasized by Shiotsu (2003), and a further study by Shiotsu and Weir (2007) makes the assertion that syntactical knowledge is relatively more important than lexical knowledge. This confirms the conclusions reached in the previous section about emphasis on word power and lexical knowledge being, of themselves, insufficient, as words in context are often more important than merely the knowledge of the word itself. LEE passages were found to closely represent FYA texts in terms of opacity of structure which, according to Berman (1984), is characteristic of academic texts. In fact, Weir et al. (2009) have advocated that, for tests of academic English to be valid, it is required that they should represent the syntactical features which are commonly found in academic texts.

Significant differences were found on G1, G2 and G5 (Table 7.2). The first two relate to G1 *Average word per sentence* and G2 *Average sentences per paragraph*. In each of these cases, unlike the significant differences found so far under vocabulary (Section 7.2.1 above), the significant difference was in the direction of LEE passages having longer sentences and paragraphs made up of more sentences than FYA texts. Also, on G5 *Mean number of words before the main verb of the main clause* measuring syntactic complexity, LEE passages were found to have a significantly higher mean number of words before the main verb of the main clause than FYA texts, signifying that LEE passages were likely to be more structurally opaque. On these 3 variables, LEE passages were not representative of FYA texts. However, academic texts are often more terse than LEE texts which often contain descriptive subordinate clauses. Excerpt[4] 1 below is a typical extract from a LEE passage which consists of only one paragraph containing 11 sentences and the average number of words per sentence is 16.545:

[4] These excerpts are exact copies of the original texts. Any errors or poorly constructed sentences have not been modified.

Excerpt1:

The first permanent settlers in Zanzibar probably arrived in about 1000 AD. Little is known about the early history of the islands, but Zanzibar became an important headquarters for merchants from Persia, Arabia, India, China, and Portugal by the year 1500. Goods such as ivory, spices, glassware, and textiles were traded *extensively*. Because of the prevalence of the spice trade, Zanzibar was known as the Spice Islands. Rich and *prosperous*, Zanzibar caught the attention of other countries. The Portuguese gained control of Zanzibar in 1504. They ruled there until 1698, when the Sultanate of Oman took power over the islands. Under Omani leadership, Zanzibar became the world's largest producer of cloves and the largest slave trading center on the east coast of Africa. In 1832, Said bin Sultan Al-Said, Sultan of Oman, relocated his capital from Muscat to Zanzibar, illustrating the importance of the islands' wealth and resources. Zanzibar became independent of Oman after the Sultan's death in 1856, and soon fell under British control. It wasn't until 1964 that the islands became part of its current country, Tanzania.

By contrast, a typical extract from FYA (Engineering) consists of much shorter and more concise paragraphs (see Excerpt 2 below):

Excerpt2:

Applications of derivative are enormous. Scientists, IT specialists, economists and engineers use the application in various branches of Science and Technology. The following are the few areas:

i **Tangent and normal to a curve**
ii Rate of change of quantities
iii Maxima and Minima

Tangent
Definition: Tangent is a line that touches a curve.

Normal
Definition: Normal is a line perpendicular to the tangent to curve at a given point.

This FYA text consists of 8 paragraphs for a total number of 10 sentences giving a mean number of sentences per paragraph of 1.25. The mean number of words of sentence is 7. Coh-Metrix is based on an algorithm which counts the number of hard returns in a passage as indicators of the end of a paragraph. This even applies to headings if the passage contains headings.

This definition of paragraph is, therefore, followed in this study as it is based on how Coh-Metrix measures a paragraph (see Chapter 3 Section 3.5).

Even in business studies, where more explanation is required, the paragraphs still tend to be shorter and more concise as in (Excerpt 3) where the passage consisted of 5 paragraphs and a total of 11 sentences giving a mean number of sentences per paragraphs of 2.2. The average number of words per sentence in this passage was 9.455:

Excerpt3:

2.1.1 Sole Traders

It is a business owned and operated by just one person. The owner is known as sole-proprietor or sole-trader.
According to D.W.T. Stafford

"It is the simplest form of business organization, which is owned and controlled by one man."
Sole proprietorship is the oldest form of business organization which is owned and controlled by one person. In this business, one man invests his capital himself. He is all in all in doing his business. He enjoys the whole of the profit. This form of organizations is very easy to start as there are few legal requirements. Example- A small shop owner.

As can be seen from this small sample presented in Table 7.3, the academic texts have shorter paragraphs and shorter sentences than the LEE passage:

Table 7.3 Summary of sample comparison between LEE and FYA extracts

Text type	Number of paragraphs in passage	Total number of sentences in passage	Mean number of sentences per paragraph	Mean number of words per sentence
LEE Excerpt	1	11	11	16.545
FYA Excerpt 1 (Engineering)	8	10	1.25	7
FYA Excerpt 2 (Business)	5	11	2.2	9.455

As has been explained, however, Coh-Metrix uses a standard algorithm based on the hard return symbol which results in headings and lists being counted as paragraphs. The LEE passage was similarly analysed by Coh-Metrix. One could argue against the inclusion of headings and lists in paragraph count but such a method for presenting information could have been used in the LEE passage: e.g. the countries mentioned could have been presented in a list. On that basis, it is still valid to compare LEE with FYA based on Coh-Metrix.

Although conventional wisdom might expect longer sentences and paragraphs to be more complex than shorter sentences, if the language used is quite indirect or embedded in the specific language of the subject area, shorter sentences can often be more difficult to comprehend than longer sentences (Khalifa and Weir, 2009). In longer sentences it is often the case that comprehension is aided by subordinate clauses which serve to amplify the meaning of the principal clause (Snow, 2002). LEE passages tend to have longer paragraphs and longer sentences and are, therefore, unrepresentative of the more terse paragraphs often found at academic level. These findings are consonant with the discussion relating to sub-

technical language and specific terms (V6, V8 and V9) under Vocabulary (Section 7.2.1).

The conclusion is that LEE passages were representative of Grammatical resources *Grammar* in relation to the variables measuring G3 *Noun phrase incidence (Syntactic pattern density)*, G4 *Mean number of modifiers per noun phrase (Syntactic complexity)* and G6 *Logical operators' incidence score (Lexical diversity)*. LEE passages were unrepresentative of FYA texts when measured on G1 *Average word per sentence*, G2 *Average sentences per paragraph* and G5 *Mean number of words before the main verb of the main clause (Syntactic complexity)*. A similar conclusion to that made on vocabulary in the previous section is appropriate here in that lower scores on G1, G2 and G5 represent academic or technical language which is characterised by a conciseness which is not typically found in passages in the Foundation Program (FP) tests.

Additionally, two further variables, *Cohesion* and *Grammar,* were assessed by expert judges. Cohesion did not show any significant difference between LEE passages and FYA texts. However, grammar was found to be significantly different between LEE passages and FYA texts confirming the three variables of grammar found to be significant using automated software analysis. Results for both variables are shown in Table 7.4:

Table 7.4 Grammar variables assessed by expert judges

Cohesion		**Grammar (*p* =.013)**	
FYA	LEE	FYA	LEE
FYA *more explicit* (*n* = 44 out of 87 texts)	*somewhat explicit* (*n* = 7 out of 15 passages)	*a balance of simple and compound sentences* (*n* = 40) or *mainly simple sentences* (*n* = 36) 48.76	*a balance of compound and complex sentences* (*n* = 13) 67.40

Although *Cohesion* generally falls under the category of Discourse mode, one aspect of cohesion relates more to Grammatical resources in considering how ideas are explicitly marked through references, conjunctions and connectors. The judgment of the experts on this feature was not found to differ between LEE passages and FYA texts although there was a non-significant difference in cohesion between both. This difference was only in terms of degree of explicitness between the two types of texts with judges agreeing that FYA texts tended to be explicit while LEE passages tended to be somewhat explicit. The conclusion is that in terms of explicitness, although LEE passages are not quite as explicit as FYA texts, they are nevertheless broadly representative of FYA texts. Cohesion is commented on in greater detail under Discourse mode below.

The second variable listed under Grammatical resources assessed by expert judges was that of *Grammar*, related to the degree of simplicity or complexity of sentences. A significant difference was found between LEE passages and FYA texts. However, the result was somewhat subtle. The median judgement in both LEE passages and FYA texts was 2, *a balance of simple and compound sentences*. The reason why a significant difference was found lies in the diversity of types of sentences found in FYA texts whereas practically all the sentences in LEE passages were rated as 2. In this sense, despite the median score being the same, LEE passages are not representative of the diversity of types of sentences found in FYA texts. Sometimes the difference lies in the direction of FYA texts being judged to be *mainly simple sentences*, a finding which resonates with the discussion above about the difficulty of comprehension of simple sentences which are indirect or quite terse (Grammar Section 7.2.2). In the other direction, FYA texts sometimes contain more *compound and complex sentences* than was found in the case of LEE passages which raises questions about the test takers' ability to be able to comprehend the more complex forms in some FYA texts. Guidelines for the degree of complexity of sentences are generally lacking in test specifications. This deficit in test specifications has been noted by (Khalifa and Weir, 2009 and Green, 2011) who have pointed out the lack of such guidelines in relation to the CEFR. This lack of clear guidance can be seen in

the diversity of practice in different testing schemes as suggested by Alderson et al. (2004) and presented in Table 7.5:

Table 7.5 Suggested guidance for sentence type

Suggested guidance	Test
Only simple sentences mapped to	Key English Test (KET)
Mostly simple sentences mapped to	Preliminary English Test (PET)
Frequent compound sentences mapped to	First Certificate in English (FCE)
Many complex sentences mapped to	Cambridge Advanced English (CAE) & Certificate of Proficiency in English (CPE)

Adapted from Alderson et al. (2004)

The scheme proposed by Alderson et al. (2004) which recommended that tests should reflect a range of sentences from mainly simple sentences through to mainly complex sentences was followed in designing the Likert scale used in this research. Practically all LEE passages were adjudged to be *a balance of simple and compound sentences*, whereas FYA texts tended to reflect the diversity of simplicity/complexity of sentences recommended by Alderson et al. Therefore, it is clear that there is a lack of guidance for test designers in Oman in terms of the representation of the different types of sentences from simple to complex. Thus, LEE passages are unrepresentative of the diversity of sentence type found in FYA texts. This is considered to be a serious deficit as Alderson et al.'s (2004) scheme reflects diversity of types of sentences necessary for valid testing of different proficiency levels.

7.2.3 Grammatical resources: Readability

The third subsection of Grammatical resources was *Readability*. Three different measures of readability were used (refer to Chapter 6 Section

6.3.3). In the case of two of these, R1 and R3, no significant difference was found between LEE passages and FYA texts (see Table 7.6):

Table 7.6 Readability variables assessed by automated analysis software

Variables with NO significant difference	Variables with significant difference
R1 Flesch Reading ease 0-100	R2 Flesch Kincaid Grade level 0-16
R3 Coh-Metrix L2 readability	

Based on R1 *Flesch Reading ease 0-100* and R3 *Coh Metrix L2 readability* measures, LEE passages were found to be representative of FYA texts. However, on R1, only approximately 25% of the FYA texts were found to be at college level (below 50) (refer to Chapter 6 boxplot Figure 6.2). This meant that the majority of FYA texts did not reach the threshold where college level skills would be required. However, none of the LEE passages reached this threshold. So, while LEE passages were representative of FYA texts, they did not reach the threshold where academic reading skills would be required. However, Khalifa and Weir (2009) used Flesch Kincaid reading ease scale to rate samples of texts from a number of international English reading texts and their findings are shown in Table 7.7:

Table 7.7 R1 Flesch Reading ease scores for Cambridge main suite level

Main Suite Level	Flesch reading ease score
KET (A2)	78.3
PET (B1)	64.7
FCE (B2)	66.5
CAE (C1)	58.4
CPE (C2)	57.7

Adopted from Khalifa and Weir (2009, p.122)

As can be seen from this table, none of these tests were found to reach the threshold at or below 50 and would therefore be deemed not to be sufficiently challenging in terms of R1.

The implication of the findings of this study is that the majority of LEE passages and FYA texts are not as demanding for the reader as similar texts found in L1 context at academic level in terms of relative numbers of syllables, words and sentences found in academic texts. It must be borne in mind that the R1 measure was designed for the US context. In the interest of fairness, at least at an early stage for L2 readers in Oman, the threshold could be set higher (less demanding) but gradually converging towards 50 as the learners progressed academically. If, for example, 60 were considered more appropriate for the early stages of academic reading ability, over half of the academic texts would satisfy this criterion and almost all of the LEE texts would be quite close to that threshold (see Chapter 6 boxplot Figure 6.2). Moreover, they would be approximately at the levels of Cambridge Advanced English (CAE) and Certificate of Proficiency in English (CPE) in Table 7.5 above. Both of these are mapped on the CEFR to levels representing proficient users of English (Levels C1 and C2) and were assigned scores on the Flesch Reading ease scale quite close to almost all the scores on LEE passages and FYA texts in this study.

An alternative to R1 and R2 as measures of readability is R3 *Coh-Metrix L2 readability*. This was included but, similar to the findings on R1, no significant differences between LEE passages and FYA texts were found. However, again similar to what was found for R1, there was a greater spread in the R3 data for FYA texts than for LEE passages as was shown by the Boxplot (Figure 6.3 Chapter 6). Although R3 represents a significant advance on R1 and R2 as measures of reading ability due to the fact that it has been possible to be able to embrace features such as semantic lexicons and syntactic parsers, results were still found to be similar to that of R1 leading to the conclusion that LEE passages were representative of FYA texts in terms of *Coh-Metrix L2 readability*.

However, the findings for R2 *Flesch-Kincaid Grade level* revealed that FYA texts were significantly lower by approximately 1.6 grades than LEE passages. Despite the significant difference, LEE passages are nevertheless still representative of FYA texts as Flesch-Kincaid Grade level takes into account the number of syllables, words and sentences in a passage. The average difference of 1.6 grades between LEE and FYA must be seen in the context that both of them are 3 to 4 grades below the threshold (13 to 16 college level US). These results resonate quite closely with those found by Weir et al. (2006) for CAE (C1) and CPE (C2) which also fell below the threshold on the Flesch-Kincaid Grade level.

At this point, it is worthwhile considering studies which validated R3 in particular since it is stated as an advance on R1 and R2 especially by the inclusion of other features such as textual processes and cohesion. It follows that since no significant differences were found on R3, that R1 and R2 cannot show any meaningful significant differences either. Therefore, on measures of readability, it is concluded that LEE passages are demonstrably representative of FYA texts in terms of R1, R2 and R3. These findings are in line with those of Weir et al., (2009) and Green et al., (2010) where no significant differences were reported between IELTS and undergraduate texts for R1, R2 and R3. While these three measures of readability are generally acceptable among some scholars, others such as Masi (2002) found that they were inadequate for revealing text complexity and difficulty. For that reason, it is important to acknowledge the fact that other factors such as semantic and syntactic complexity should also be taken into account (Weir et al., 2009). It follows from this that, in this study, a more holistic approach is adopted by taking into consideration results from other measures (such as Grammatical resources: Vocabulary and Grammatical resources: Grammar), which embraced quantitative and qualitative measures. Additionally, R1, R2 and R3 results were not considered in isolation from results obtained by other measures in order to add depth to or to critique these readability measures.

7.3 Discourse mode

Discourse mode refers to how the text is organised for the purposes of comprehension. A number of measures of cohesion were examined (see Chapter 6 Section 6.3.4) and these are set out in Table 7.8:

Table 7.8 Discourse mode: *Cohesion* variables assessed by automated software analysis

Variables with NO significant difference	Variables with significant difference
C1 *Anaphor reference*	C4 *Latent semantic analysis (LSA)* is a major of semantic or conceptual similarity between text excerpts
C2 *Argument overlap*	C5 LSA sentences all combinations mean
C3 *Proportion of content words that overlap between adjacent sentences*	

These five variables were assessed quantitatively using automated software. A further four variables were assessed qualitatively by expert judges' opinions and these are presented in Table 7.9:

Table 7.9 Discourse mode variables assessed by expert judges

Variables with NO significant difference	Variables with significant difference
Rhetorical task	Genre (p = .020)
Pattern of exposition	Rhetorical organization (p = .006)

The findings related to these variables are now discussed in the following section.

7.3.1 Cohesion

No significant differences between LEE passages and FYA texts were found in terms of the first 2 variables, C1 *Anaphor reference* and C2 *Argument overlap*. In terms of C1, which measures the degree of backward reference up to five sentences earlier in the text as a measure of cohesion in a text, LEE passages were found to be representative of FYA texts. Similarly, in the case of C2 which measures the degree of noun, pronoun or noun phrase in one sentence being a co-referent of a noun pronoun or noun phrase in another sentence (Green et al. 2010), LEE passages were found to be representative of FYA texts. In either case, the Coh-Metrix scores showed moderate degrees of cohesion (C1 FYA *M* = .42, LEE *M* = .33; C2 *M* = 0.35*, M* = 0.34). Essentially these results indicated passages which, to a moderate degree, aided comprehension by these cohesive devices. However, the conclusion is that LEE passages reflected the degree of cohesion found in FYA texts. Higher scores on C1 and C2 would tend to reflect more difficult texts (McNamara et al., 2012) than would be expected at foundation program exit level or first year academic level.

Similarly, in terms of C3 *Content word overlap*, no significant difference was found between LEE passages and FYA texts. The measure of content overlap takes into account the number of words in a sentence. Thus, longer sentences would result in lower scores on this variable (McNamara et al., 2012). Both LEE passages and FYA texts tend to have longer sentences than the more terse texts found in advanced academic texts. This resonates with the results found for V2 *Standardised type-token ratio* and V3 *Lexical density* where, although no significant differences were found, the scores on those variables were determined by the overall number of words in the sentence or text. For example, 1 content word in a 5-word sentence would score higher than 1 content word in 10-word sentence. There is thus some concordance between C1, C2, and C3 on one hand and, V2 and V3 on the other, considering that scores on both are based on sentence length.

However, on measures of conceptual cohesion (C4 and C5), LEE passages were found to be significantly less cohesive than FYA texts. Both C4 and C5 measure cohesion but the former is more local in scope basing its measure

on similarity of meaning in adjacent sentences whereas the latter is based on similarities of meaning between all sentences within a text as a whole. Similar results were found by Green et al. (2010) although only on C5 was a significant difference found. An explanation for this difference between LEE type passages (IELTS in the case of Green et al.'s (2010) study) and FYA texts is alluded to by the authors as being attributable to the need for LEE type passages to provide rich information in a relatively short passage. This was required in order to enable test questions to be framed, whereas, by contrast, academic texts tend to give attention to a topic of more limited scope (Green et al., 2010) and, consequently, tend to be quite terse. This means that LEE passages were not representative of FYA texts based on measures of conceptual cohesion. The difference found between LEE passages and FYA texts was, however, simply a matter of degree. A plausible explanation for this difference lies in the fact that it is expected that as students progressed academically their texts would also tend to exhibit greater degrees of cohesion. Thus, average scores would be expected to be higher than this once their textbooks became more specialised in nature, as the students would tend to have a greater command of the specialist vocabulary and concepts involved.

The discussion relating to the measures of global cohesion resonates with the earlier discussion related to Grammatical resources: Vocabulary (Section 7.2.1) and also to Grammatical resources: Cohesion (Section 7.2.2). Under Vocabulary, LEE passages were found to be significantly lower in their use of specific terms and technical vocabulary than FYA texts. But even more to the point, Cohesion viewed from the perspective of how ideas in a text are explicitly marked through reference, conjunctions and connectors, the findings of the expert judges revealed a significant difference between LEE passages and FYA texts. However, this difference was only in terms of degree of specificity, FYA texts being judged to be *specific* while LEE passages were judged to be *somewhat specific.* All three findings taken together lead to a more consistent conclusion in that the findings resulting from automated software have been supported by the more qualitative ratings of the expert judges.

On measures of *Rhetorical task* and *Pattern of exposition* there were no significant differences found between LEE passages and FYA texts. These two variables were scored by taking into account the decisions of expert judges (Chapter 6 Section 6.4.2 and Section 6.4.6).

In terms of *Rhetorical task* assessed by the judges' decisions as to the category of text which best described the passages, for both LEE and FYA there was general agreement that the texts mostly fell into the category of *exposition*. Although many authors agree that the readers' understanding of how the text is organised greatly influences their reading comprehension (e.g. Anderson 1999b; Farrall, 2012), there is less unanimity on the classification schemes for these rhetorical tasks. In this study, Weir et al (2009) were followed in their adaptation of Enright et al. (2000) in categorising the rhetorical tasks under three headings. These were *exposition*, *argumentation/persuasion/evaluation*, and *historical biographical narrative*. That the judges should decide that in the case of both LEE and FYA the texts were predominantly of the exposition type is hardly surprising in view of the fact that most college texts tend to be informative in the sense of transmitting knowledge (Enright et al., 2000). The findings for rhetorical task were very similar to the result of Weir et al.'s (2009) and Green et al.'s (2010) studies where most of the texts were found to be expository in nature for both IELTS and undergraduate texts and hence, LEE passages were found to be representative of FYA texts.

The second of these variables, *Pattern of exposition*, revealed no significant difference between the judges' decision as to the type of exposition that was involved. However, more of the LEE passages tended towards an *elaborate* pattern than was the case for FYA texts, but in both cases many of the passages were classified as the *describe* pattern of exposition. It is somewhat surprising that so few texts, especially FYA texts, were classified under such patterns: *cause-effect* or *problem-solution* or *compare-contrast,* patterns that would be expected to be found in more academic type texts. A possible explanation is that academic texts examined by the expert judges tended to be taken from more elementary text books and these types of texts are often quite descriptive. Patterns such as *problem-solution* tend to characterise reading for integrating information in academic settings (Grabe,

2009). These results are in contrast with the findings of the Weir et al. (2009) study where the IELTS texts tended to exhibit more instances of problem-solution and cause-effect. It must be remembered, however, that the IELTS test tends to target students seeking entry into universities in English speaking countries. These require students to achieve at least band 6 in IELTS for entry purposes. By contrast, the threshold for entry to the Colleges of Technology in Oman tends to be 5 (OAAA, 2008). Also, these authors found that the judges tended to identify the pattern of exposition of university undergraduate texts as more elaborative whereas, in this study, it was the LEE passages which were adjudged to be more elaborative. However, it has to be admitted that the decisions of the judges can often be subjective and in practice it might have been difficult for them to make a fine distinction between descriptive and elaborative texts, for instance.

Weir et al. (2009) have pointed out the difficulty of selecting a short passage from a lengthy piece of text at academic level in such a way that it could be free standing and not be dependent on earlier paragraphs for comprehension. This same difficulty was encountered in this study as many passages from academic texts had to be rejected because they would not make sense taken out of the overall context. It was difficult to find suitable extracts that could be used for the purposes of this research. Thus, the passages selected may have been more representative of descriptive or elaborative patterns rather than passages exhibiting cause-effect or problem-solution as short passages of these types often depended on earlier paragraphs for comprehension.

The final two variables assessed by expert judges were *Genre* and *Rhetorical organisation*. On both of these variables significant differences were found between LEE and FYA. Genre was assessed by classifying texts across a range of genre types. LEE passages were significantly more often identified as *magazines or newspaper articles* type whereas FYA texts were more often described as *text book* type. This significant difference is nonetheless an unsurprising result since all of the FYA texts were extracted from college text books and would therefore exhibit the characteristics of such a genre, whereas LEE passages were frequently extracts from

magazines or newspapers or articles and would be expected to exhibit characteristics typical of that genre. Although a significant difference was found for genre, other genre types were found to be represented in both LEE passages and FYA texts. However, these findings lead to the conclusion that on genre, LEE passages were not representative of FYA texts. These results were similar to the findings of Weir et al. (2009) where IELTS texts tended to be of the newspaper/magazine journalistic type genre and the academic texts were predominantly textbook type. Green et al.'s (2010) findings were similar in that the judges identified newspapers/magazines articles as the main genre for IELTS passages although they found greater diversity in undergraduate text genres. Because the findings of this study revealed a significant difference, it is concluded that, in terms of genre, LEE passages were unrepresentative of FYA texts, a finding supported by Moore et al. (2011) who also found that IELTS passages did not reflect the more pragmatic and critical genres found in academic study.

The last feature examined by the expert judges was *Rhetorical organisation* which referred to their decisions as to how explicit the organisation of the text was. A significant difference was found between LEE passages and FYA texts. Most FYA texts were adjudged to be *explicit* while most LEE texts were adjudged to be *somewhat explicit*. The difference, therefore, was in terms of degree of explicitness. Nevertheless, on rhetorical organisation, LEE passages were not representative of FYA texts. A similar result was found for *Cohesion* measured in terms of Grammatical resources (Section 7.7.2). Cohesion was measured in terms of how explicitly ideas were linked through references, conjunctions and connectives. LEE passages were significantly less explicit than FYA texts but again this was a question of degree of explicitness where LEE passages were adjudged to be *somewhat explicit* while FYA texts were adjudged to be *explicit*. This result is at variance with the studies by Green et al. (2010) and Weir et al. (2009) where no significant difference was found between IELTS and undergraduate texts on rhetorical organisation.

The results for rhetorical organisation resonate with the earlier result found for pattern of exposition where FYA texts were found to be predominantly

descriptive but LEE passages were found to be *descriptive* and *elaborative*. The implication is that LEE passages are actually more challenging due to the inclusion of more elaborative texts than FYA texts which were found to be predominantly descriptive. Actually, Khalifa and Weir (2009) have argued this very point based on evidence from a number of studies in which it was found that comprehension was aided to a greater extent by problem-solution, causation or comparison mode than descriptive or other types. This subtle difference in terms of the degree of explicitness has made it, at least to some extent, more difficult for LEE passages which were found to be only somewhat explicit. This is because of the organising frames which render the exact meaning of the text to some degree more difficult than it would be for a more explicit rhetorically organised text.

7.4 Nature of information: Abstractness

One of the contextual features of a text is the nature of the information it conveys in terms of its abstractness or concreteness (see Chapter 2 Section 2.7.3.7 and Chapter 6 Section 6.3.5). Khalifa and Weir (2009) have commented on how more concrete words tended to draw on a range of cognitive processes such as associated imagery whereas abstract words tended to rely more on verbal systems alone. It is often assumed that academic texts are characterised by more abstract words and so it was important to test this assumption by examining LEE passages and FYA texts to measure their concreteness-abstractness.

The abstractness of the texts was measured by four variables as presented in Table 7.10:

Table 7.10 Abstractness variables assessed by automated software analysis

Variables with NO significant difference	Variables with significant difference
A1 *Concreteness for content words, mean*	A3 *Mean hypernym value of nouns*
A2 *Concreteness minimum in sentence for content words, mean*	
A4 *Mean hypernym value of verbs*	

The first three of these revealed no significant differences between LEE passages and FYA texts leading to the conclusion that the test passages were representative of the academic texts used in the first year in relation to their abstractness. A1 is an indicator of the average content words in a text. Low scores would indicate words tending to be more abstract. In the analysis chapter (Chapter 6 Section 6.3.5), FYA texts were found to be slightly more abstract than LEE passages but this difference was not significant. Returning to the actual scores for A1, for LEE passages the score (*M* = 374.4) indicated an average of 374.4 words out of passages of average length of 500. This equates to approximately 75% of the words being more concrete. By comparison and taking into account no significant difference, it is inferred that FYA texts would also average approximately 75% for concrete words in the overall texts. Thus, both types of texts are quite high in concreteness of content words, despite the earlier assumption that there should be a higher content of academic (more abstract) words at this level. In fact, Green and Hawkey's (2011) study has challenged this assumption by making a plea for IELTS passages to be more straightforward (see Chapter 2 Section 2.7.2.7). The assumptions that LEE passages and, *a fortiori*, FYA texts should rate higher in terms of abstractness is based on a slight misnomer in the application of the term 'academic'. Although in the Omani context, first year college programs are located at academic (third) level, actually the subjects to be studied are more technological or vocational in nature (see Chapter 1 Section 1.1 and Section 1.2). So, although these subject areas do require an academic or theoretical underpinning, their primary impetus is that of the

application of knowledge in practical contexts. More advanced learners tend to use more abstract language (Coady's (1979) model of ESL learners) but this generally occurs beyond first year academic. The LEE passages are aimed at assessing the students' ability to be able to benefit from study at first year academic. For this reason, the conclusions of Green and Hawkey (2011) that proficiency test items such as those found in IELTS, that are more straightforward, are also more apposite in assessing a student's readiness for reading at FYA. Thus, the finding that LEE passages were representative of those found at FYA texts is also supported in the literature which recommends the use of more straightforward passages in proficiency tests.

However, in a similar study though in a different context, Green et al. (2010) reached different conclusions finding that academic texts were significantly more abstract than the IELTS passages. Given that the average IELTS passage in their study was 854 words, this meant that on average 44% of the words were found to be concrete compared with 75% in this study. This explains the plea by Green and Hawkey (2011) for more straightforward IELTS passages. In the context of the current study, this indicates a good fit between the LEE passages and FYA texts. The difference in results on A1 between the current study and Green et al.'s (2010) may lie in the different contexts of each study. Green et al. (2010) were investigating texts intended for both L1 and L2 readers in the UK context. Secondly, the academic texts were extracted from a range of more academic subjects such as sociology where students would be expected to encounter more discursive and nuanced language. The context of the current study differs firstly in terms of target learners who were all L2 learners in a non-English speaking context and secondly the range of subjects at academic level which, with a few exceptions, were mostly technological in nature (see Chapter 2 Section 2.7.3.6 and Chapter 1 Section 1.1 and Section 1.2).

A2 focuses on each sentence in a text and assigns the lowest score for concreteness that has been identified. For a passage the mean of all these scores is then taken across all the sentences as an indicator of abstractness. The lower the score, the more abstract the passage tends to be. There was a

very small difference, which proved to be non-significant between LEE passages and FYA texts, with the latter scoring being just marginally lower and, therefore, more abstract than the former.

A4 assesses the abstractness of verbs in a text by means of their hypernym values i.e. a measurement based on a conceptual hierarchy in which the number of levels of subordinate words below and superordinate words above the target word indicate the relative concreteness or abstractness of that word (McNamara et al., 2012; McNamara et al., 2005). The lower the value (fewer levels above the target word) the less specific (concrete) the word is (McNamara et al., 2012). Only a small non-significant difference in scores was found for LEE passages and FYA texts. A similar result was reported in Green et al. (2010) who found that for A2 (Concreteness minimum in sentence for content words, mean) and A4 (Mean hypernym value of verbs) there were no significant differences between IELTS passages and academic texts. Thus, in terms of these three variables (A1, A2 and A4) LEE passages were found to be representative of FYA texts.

However, on A3, the *Mean hypernym value of nouns* in the text, a significant difference was found between LEE passages and FYA texts. FYA texts had a significantly higher hypernym value of nouns (M = 6.38) compared with LEE passages (M = 5.47). The academic texts were found to contain more specific terms than the LEE passages. Although specificity is not co-terminal with abstractness-concreteness, it is expected that less specific terms would tend to be more abstract. Although a significant difference on A3 is not supported by the Green et al. (2010) study, it is consistent with findings elsewhere in this study, for example the findings for Grammatical resources vocabulary (V6, V8 and V9 in Section 7.2.1) where a significant difference was found in terms of specificity with more specific and technical terms characterising academic texts. Another result, which strengthens the consistency of this current study, was the significant difference which was found in cohesion where FYA texts displayed stronger degrees of specificity than LEE passages.

7.5 Functional resources

Functional resources refer to the illocutionary force of a passage or part of a passage in which the meaning does not simply reside in the literal sense of what is written (see Chapter 2 Section 2.7.3.4). The conveyance of implied meanings takes place through a number of devices identified in the literature. These are *ideational (*Descriptions, classifications, explanations…etc.), *manipulative (*Requests, suggestions, commands, and warnings…etc.), *heuristic* (for teaching and learning, problem solving, retention of information) and *imaginative* (jokes, and use of figurative language and poetry).
Passages from LEE and FYA texts were assessed by the expert judges and a significant difference was found (see Chapter 6 Section 6.4.3). Both LEE passages and FYA texts were mainly identified as either *heuristic or ideational*, although a small number of FYA texts were also assessed as *manipulative and imaginative*. Whereas LEE passages were mostly ideational (*n* = 9, 60%), FYA texts were found to be mainly heuristic in nature (*n* = 61, 71%). Thus, LEE passages were unrepresentative of the functional resources that mainly characterised FYA texts.
These results resonate with what was found for measures of concreteness-abstractness under A3. FYA texts were found to have significantly more specific words and more abstract words than LEE passages. Since foundation students have not yet progressed to learning the principles of the various sciences, the passages used for their tests tended to be more general and descriptive in nature. Thus, they made more use of ideational devices which are characteristic of descriptive language. One example of this from the foundation passages referred to the shift of the capital city from Muscat to Zanzibar. On a purely literal level this would mean the moving of the entire city, brick by brick, to a new location but the ideational force of this language was a descriptive way of signifying a shift of power. For academic texts however, underlying principles are being introduced. These require strict definitions or the memorisation of formulae which become devices for *solving many kinds of problems*. Texts from business studies often exemplify some principle by the use of a case study. Engineering principles are often stated in symbolic formulae. So, although LEE passages are not

representative of heuristic devices, this is the result of the change in the nature of what is being read at foundation level when compared with what is read at the academic level. A glance at the curriculum in FYA shows that the subjects to be studied are often stated as 'principles' of accounting, management, etc. Learning the rudimentary laws and principles of a subject necessarily requires heuristic devices which may be stated as principles, theories, laws, hypotheses and formulae. Typical extracts used for LEE passages often consisted of magazine or newspaper articles which did not report by citing in-depth principles or definitions as would be the case with academic textbooks. This resonates with the results found for genre where the expert judges assessed LEE passages as predominantly magazine/newspaper type, significantly different from their assessment of FYA texts which were classified mostly as research or academic type (Section 7.3).

7.6 Content knowledge

Content knowledge refers to the extent to which the comprehension of a passage is dependent on the background knowledge of the reader (see Chapter 2 Section 2.7.3.8). It is desirable that passages used for English test purposes should not be overly reliant on background knowledge for comprehension as this might give rise to issues of validity and fairness (see Chapter 2 Section 2.7.3.8).

Content knowledge in this study was examined under four headings (see Chapter 6 Section 6.4.5) as presented in Table 7.11:

Table 7.11 Content knowledge variables assessed by expert judges

Variables with NO significant difference	Variables with significant difference
Cultural background Language background Religion background	Subject specificity

Only in the case of subject specificity was a significant difference found between LEE passages and FYA texts.

Thus, on the other three variables, namely the influences of cultural, language and religion backgrounds on comprehension, LEE passages were representative of FYA texts. This result was salutary in that academic texts tend to be less reliant on these three variables for comprehension and that these variables were not over-represented in the LEE passages.

In relation to subject specificity in content knowledge, a significant difference was found. Almost all of the judges' decisions identified the LEE passages as either *general* in nature (53%) or else *neutral*. In contrast, there was a greater variety in their classifications of the FYA texts but with the majority being classified as *specific* (59%). This result is what might be expected intuitively, as subject specific content knowledge as a requirement for comprehension would be more characteristic of technological subjects whereas LEE passages would reflect extracts from newspapers or magazines which would be less reliant on subject specific knowledge. This result corresponds with the findings of both the Green et al. (2010) and the Weir et al. (2009) studies although both of these studies also found a significant difference for cultural background. Thus, LEE passages were found to be unrepresentative of FYA passages in terms of subject specificity but were found to be representative in terms of the other three variables indicating that the LEE passages were not overly reliant on cultural, language or religion content knowledge.

7.7 Overall text purpose

Overall text purpose refers to the extent to which the knowledge of the purpose for which the text has been written aids comprehension. As identified in the literature (Chapter 2 Section 2.1), this can be assessed by means of considering five different types namely, *referential* (intended to inform), *conative* (intended to persuade), *emotive* (intended to convey feelings or emotions), *poetic* (intended to entertain, delight, please) and *phatic* (intended to keep in touch).

No significant difference was found between LEE passages and FYA texts with almost all the passages being classified by the judges as *referential* (FYA 90% and LEE 93%) (see Chapter 6 Section 6.4.6). Thus, the informative nature of typical LEE passages' overall text purpose were found to be representative of the referential nature of the overall text purpose of the FYA texts.

This is a pleasing result in terms of the representativeness of the test passages. However, even in technological subjects, some understanding of other types of purposes would be advantageous. For example, conative and emotive purposes may be found in marketing strategies (say for Business Studies), and conative purposes may be found in legal aspects of some courses. As students progress further they need to be able to distinguish between texts based on fact and evidence from those which are largely based on opinion. For example, in Key English Test (KET), texts were found to be mainly referential in terms of overall purpose but for more advanced tests such as Cambridge Advanced English (CAE) and Certificate of Proficiency in English (CPE), texts represented a variety of purposes including referential, poetic, emotive and conative (Khalifa and Weir, 2009). Moreover, the recent development of more collaborative approaches to learning requires students to communicate through chat rooms and blogs where contributions might not be so referential in terms of their purposes. Thus, an ability to recognize the nature of other purposes in texts could become more important. Indeed, Khalifa and Weir (2009) have intimated the importance of more phatic and emotive writing in view of the growing use and importance of social media.

This result resonates with the earlier findings for discourse mode where *rhetorical task* was found to be of an *expository* type for both LEE passages and FYA texts and, again, for *pattern of exposition* where both LEE passages and FYA texts where found to be *descriptive* in nature.

7.8 Writer-reader relationship

This feature recognizes the importance of the text as an intermediary between the writer and the reader (see Chapter 2 Section 2.7.3.2). Ede and

Lunsford (1984) made a distinction between two types of audience, the audience addressed and the audience invoked. The latter refers to a more fictional audience being addressed for rhetorical purposes. No significant difference was found between LEE passages and FYA texts as the judges assessed that over 90% of the texts of both LEE and FYA texts were directed towards the *audience addressed* (see Chapter 6 Section 6.4.6). This result would be expected, as technological passages would be directly informative rather than rhetorical, as was found for the referential nature of texts earlier in this study (Section 7.7). Thus, LEE passages were representative of FYA passages in terms of writer-reader relationship. However, the implications of writer-reader relationship points to the difficult task confronting test designers who select passages based on their assumptions of the extent of the test taker's knowledge on various subjects. The corollary of this is that test designers should have an awareness of the types of students being tested and this was likely to have been the case in terms of the Omani students.

7.9 Task setting

7.9.1 Channel of presentation

Channel of presentation refers to characteristics of tasks whereby comprehension is aided by other features, for example, diagrams or images (refer to Chapter 2 Section 2.7.2.5). A significant difference was found between LEE passages and FYA texts but this difference was in terms of degree of appropriateness (see Chapter 6 Section 6.5.1). The judges' assessments tended to rate test passages as *somewhat appropriate* whereas they tended to rate academic texts as *appropriate*. It is therefore concluded that LEE passages are not representative of FYA in terms of channel of presentation. A prima facie view of the texts showed that almost all of the test tasks used in the study had passages that were not accompanied by other features such as graphs, charts or even headings. There was only one case where the passage contained a table. By contrast, almost all of the academic texts included additional features such as charts

and images. As discussed in the literature, comprehension is greatly aided by the presentation of information in more than one channel (see Chapter 2 Section 2.7.2.5). For that reason many proficiency tests include passages which contain diagrams or charts in addition to text, for example, the Canadian Academic English Language (CAEL) test which incorporates a task involving the labeling of a diagram. Test tasks in the Omani context are, therefore, more difficult to read and comprehend due to being generally restricted to one channel. Even though these test passages were largely descriptive in nature, comprehension would have been considerably augmented by the inclusion of maps, charts or tables.

7.9.2 Text length

A difference was found between LEE passages and FYA texts in terms of the expert judges' ratings on a five point *Likert* scale ranging from *appropriate* to *not appropriate* (see Chapter 6 Section 6.5.2). The median score for FYA texts indicated that they were rated as appropriate whereas LEE passages were rated somewhat appropriate. However, this difference was found to be non-significant. Thus, it is concluded that LEE passages were representative of FYA texts in terms of appropriateness of passage length.

The significance of text length as a task setting lies in its use for ensuring a passage is long enough to allow scope for cognitive processing (Nuttall, 1996; Spyridakis and Standal, 1987). Additionally, the literature stresses the need for multiple passages in order to test the readers' ability to create their own organising frames, which might not be possible by relying on the use of a single text (Grabe, 2009). In the Omani situation, only a single text is used of approximately 500 words. Clearly, this is not sufficient to test comprehension through organising frames; however it is also agreed in the literature, as pointed out in (Chapter 2 Section 2.7.2.6), that there is no prescription for the correct number of texts nor for the lengths of those texts. The guiding principle is the purpose for which the test is being conducted. In international tests there is a wide variation in the number of texts used ranging from 3 to 9 in the Cambridge Main Suite reading papers (see

Chapter 2 Section 2.7.2.6). It would appear that three different types of texts would be the minimum requirement. The KET, which is the equivalent of A2 level (basic user) on the CEFR, for example, consists of 4 texts with an overall word count of approximately 740-800 words. Nevertheless, the Omani LEE passage length was representative of the FYA text lengths based on the judges' assessment. Thus, while text lengths were not at issue, the judges had not been asked to express their opinion about the number of texts required in LEE. The absence of multiple texts in the LEE may limit the opportunity for exercising discernment in terms of what to read expeditiously and what to read carefully. Such a deficit has been commented on by Weir (2005) as possibly affecting the validity of the test.

7.10. Summary

Vocabulary, grammar, readability, cohesiveness, nature of information, linguistic demands and task setting were measured by automated software or by the considered opinions of expert judges. In general, LEE texts and tasks were found to be representative of FYA texts.

Some significant differences were found between LEE and FYA but these were usually in the direction of LEE texts tending to be more general than FYA texts which tended to be more specific in nature. Evidence for this was found, for example, in the case of the third 1,000 word frequency list which contains mainly sub-technical words. FYA texts were found to contain a significantly higher number of these words than LEE passages, although no differences had been found on the first and second most frequent 1,000 words. In contrast, grammar measured in terms of words per sentence and sentence per paragraph revealed that LEE passages were actually higher than FYA texts. However, the relatively shorter sentences and paragraphs in FYA also indicated texts which were terse and more difficult to comprehend, but sometimes the reverse was true due to the higher density of logical operators in LEE passages making comprehension difficult.

LEE passages were generally representative of FYA texts on measures of readability. The one exception was R2 (Flesch Kincaid Grade level 0-16)

where a significant difference was found, but this difference had little meaning in view of the fact that both LEE and FYA fell below a threshold set in the US which had little relevance in the context of this study. Differences in terms of specificity were easiest to demonstrate in terms of genre where FYA texts were, unsurprisingly, classified by the expert judges as being of academic or research type, whereas the genre represented by LEE passages tended to be that of magazine excerpts. A similar result was found on functional resources where LEE passages tended to be ideational in nature in contrast to FYA texts which often reflected the more heuristic nature of elementary science and technology text books.

One significant difference was that LEE passages in Oman relied on the single channel of presentation of written text whereas comprehension of academic texts was generally aided by using multiple channels such as text supported by graphs, charts or drawings.

Notwithstanding this difference, the two research sub-questions (RSQ 2 & RSQ 3) have been addressed. For RSQ 2, the findings of this study led to the conclusions that LEE passages generally reflected FYA texts and differences were only found for representation of the subject specificity, genre type and heuristic nature of academic texts which in practice are difficult to replicate in Foundation Programme tests. However, on RSQ 3, while LEE tasks were representative of FYA tasks in terms of text length, they were found to be unrepresentative in terms of channel of presentation. LEE tasks were confined to a single channel, namely written texts, whereas FYA tasks used multiple channels such as written texts supported by graphs, charts or drawings.

The key findings of this chapter, which were intended to address the two research sub-questions (RSQ 2 & RSQ 3), are brought together with the key findings from (RSQ 1, Chapter 5) in order to arrive at the conclusions and implications of the research in the following chapter. Additionally, note will be made of limitations as well as indicating directions for future research.

Chapter 8 Conclusions, limitations, and recommendations

8.1 Introduction

The focus of the current research was on the testing of reading in English for Academic Purposes (EAP) in Oman. It aimed at evaluating the Level 4 Exit Exam (LEE) in the Foundation Program (FP) at the Colleges of Technology (CTs). Test results were being used to inform decisions regarding the preparedness of the test takers to undertake academic study through the medium of English. It was, therefore, important to assess how validly these test results allowed for inferences to be made about students' performances. It was of equal importance for context validity to establish how closely the texts in the reading section of the LEE resembled the texts that students would be likely to encounter at first year academic level (FYA). Essentially, this implies that the test tasks are assumed to be a sample of the wider range of tasks with which the test takers would be likely to be confronted once they advanced to academic level. In short, the aim was to ensure the context validity of the reading tests by adopting Khalifa and Weir's (2009) model. In order to understand and investigate the main issues, this study, in line with recent models of reading and testing of reading, has considered the contextual features as well as the cognitive processes by which the test takers engaged with the test tasks (see Chapter 1 Section 1.4 and Chapter 2 Section 2.3).

Three research sub-questions were posed in order to articulate further the central research question (see Chapter 3 Section 3.2). The first of these research sub-questions aimed at identifying the cognitive processes by which students engaged with the texts and test tasks. The other two sub-questions aimed at assessing how closely both the reading texts and tasks in LEE reflected those encountered in the first year of academic study.

8.2 Conclusions

8.2.1 Test taker's cognitive processes

The first research sub-question aimed at identifying the cognitive processes by which the test takers engaged with the text and tasks in the reading tests:

"What are the cognitive processes by which students engage with the texts and tasks in reading tests?"

In a natural experiment, which simulated test conditions, the test takers were asked to complete a questionnaire designed to identify the particular reading process which was involved for each task (see Chapter 3 Section 3.4 and Chapter 4 Section 4.2). Correlation tests and factor analysis indicated that there were two underlying components which accounted for the processes reported by the test takers as being involved in reading and comprehension (see Chapter 4 Section 4.3 and Section 4.4).

The first of these consisted of *basic reading* processes which were elaborated further in the discussion chapter (Chapter 5). However, in relation to the test in Oman, none of these was found to have a significant impact on test scores (see Chapter 4 Section 4.6.1). Choosing the process which was intended to be tested by the test designers had not resulted in significantly higher scores on the test than choosing an alternative process. Thus, the validity of items designed to test the cognitive processes on Component 1 are in doubt. The main cognitive processes on Component 1 included *scanning expeditiously*, *careful reading global* and the knowledge and use of *grammatical resources* for comprehension. It is therefore concluded that although these were important cognitive processes in reading and comprehension, in the case of the Omani test takers, these were still underdeveloped in terms of what would be required of them for reading at academic level (see Chapter 4 Section 4.6.1 and Chapter 5 Section 5.2).

The second component consisted of processes of expeditious reading which included *expeditious search reading*, *the use of discernment* and the use of instructions in the *rubric* for comprehension (see Chapter 4 Section 4.4 and Section 4.5). On Component 2, evidence was found that those who chose

the intended strategy also scored significantly higher on the test results (see Chapter 4 Section 4.6.2). Thus, the test items designed to measure these cognitive processes were doing so validly and so, there is confidence in the inferences made on the basis of the cognitive processes in Component 2 (expeditious reading) by which the students engaged with the texts and tasks in the reading tests (see Chapter 4 Section 4.6 and Chapter 5 Section 5.3).

8.2.1.1 Component 1: Basic reading processes

Scanning expeditiously involves the location of specific items of information within a sentence or phrase. It is an important skill for academic reading as students often have a heavy reading load and need to be able to locate required information quickly and expeditiously (see Chapter 2 Section 2.3, Chapter 4 Section 4.61 and Chapter 5 Section 5.2.1). Although 79% of the test takers reported that they had used expeditious scanning, the remaining test takers were still as successful in answering the questions correctly by means of an alternative process (see Chapter 4 Section 4.6.1.1). Not only does this fact raise issues of validity surrounding the test items as discussed in Chapter 5, Section 5.2.1, but it also draws attention to the test conditions which allowed a significant number of test takers the time to be able to read the passage more carefully in order to answer the question. Alderson (2000) has insisted on the need for comprehension within the context of scanning expeditiously to be a strictly time-bound exercise in order to more closely resemble scenarios in the real word or in the university. Accordingly, time constraint is one of the context validity features of Khalifa and Weir's (2009) socio-cognitive model adopted in this study.

In the case of the reading test which was the focus of this study, there was no time control for these items. Accordingly, these test items cannot be assumed to validly measure scanning expeditiously. Even in the case of the 79% who reported that they had used expeditious scanning, it is likely that more time was allocated to answering these test items than was justified in terms of the level of comprehension that was required. Due to the L1-L2 transfer effect, commented on by Palmer et al. (2007) (see Chapter 5

Section 5.2.1), it is likely that, when L1 is a high salience language such as Arabic is, test takers would be most likely to pay great attention to individual words rather than taking into account the meaning of a group of words. The processes involved in reading and comprehension of Arabic involves focusing on individual words whereas scanning expeditiously is a skill of looking at a group of words for specific points. This would explain why scanning expeditiously did not load significantly on Component 2 which consisted of expeditious reading processes but, instead, loaded on Component 1 which consisted of basic processes. This highlights the underdeveloped nature of scanning expeditiously among Omani test takers due to their tendency to fixate on individual words rather than attempting to discover the meaning of a group of words.

However, despite the importance attached to scanning expeditiously for students at academic level faced with an extensive reading list, Weir (2013) has commented on its relatively neglected status in the US and the UK. In more developed countries, approaches to teaching and learning of reading and comprehension have tended to focus on careful reading. Thus, it is not only in a Middle Eastern context, but more generally, that greater attention needs to be paid in learning and testing of reading to the development of expeditious reading processes including scanning at local level.

Another reason for not adopting scanning expeditiously may be due to the test takers' lack of rapid word recognition, which would have adversely affected their comprehension (Yoshimura, 2000; Koda, 2005). Their restricted word recognition skills would have made a properly time-bound comprehension test item into a formidable task. Furthermore, the test takers appeared not to have advanced sufficiently to make use of abstract words in comprehension and this, at least partially, explains why scanning expeditiously loaded on Component 1 (basic processes) and not on Component 2 (expeditious reading). In fact, it seems likely that the test takers drew on their prior knowledge (a basic reading process, see Chapter 5 Section 5.2) rather than scanning the passage to locate the required information.

In practice however, it is also difficult to measure the skill of scanning expeditiously especially by the use of binary MCQ where the test taker has a choice between true/false responses or yes/no responses (see Chapter 2 Section 2.1 and Chapter 5 Section 5.2.1). Moreover, where test items repeat verbatim what is in the passage and, worse still, where the items are presented in exactly the same logical or chronological order as in the passage used in the Omani test, validity is very much in doubt.

In careful local reading, the explicitly stated main points within a sentence are accurately established (Weir et al., 2009). According to the description of careful local reading in Khalifa and Weir's model (2009), which was adopted in this study along with the context validity from Weir's (2005) validation framework, establishing accurate comprehension usually demands the use of both lexis and syntax. 65% of the test takers did choose the intended process for answering the related test items. The remaining 35% reported using some other process (see Chapter 4 Section 4.6.1.2 and Chapter 5 Section 5.2.2). However, there was no significant difference between these groups in terms of answering the test items correctly. Because these test items do not allow for a distinction to be made between those using the intended process and those using some other process, the validity of the test items are therefore in doubt and also the validity of inferences made on the basis of these test scores regarding the students' readiness for academic reading using this process (see Chapter 5 Section 5.2.2). As discussed in Chapter 5 Section 5.2.2, there were only two alternatives to using the intended process; either the test takers used guesswork in a three way multiple choice situation or else the word 'prosperous' was already part of their vocabulary. That test takers could fare just as well by guesswork using careful reading at local level calls into question the validity of the test items purporting to measure this feature in the Omani test.

With reference to lexis, Cooper (1984) has drawn attention to the difficulty faced by L2 readers whose secondary education had not been through the medium of the second language. It was found that those whose secondary education had been through the medium of the intended second language had a much greater vocabulary level when compared with those whose

secondary education had been through L1. In the case of the test takers in this research, English had been taught in secondary school as a discrete subject and the remaining subjects were taught through Arabic (see Chapter 1 Section 1.1). Their lexical deficit and poor understanding of syntax as a result, would have presented obstacles to their being able to effectively apply careful reading.

The corollary is also true. The fact that those who reported using the intended process had not fared any better than those who guessed the answer also implies that they were not applying careful local reading with any degree of efficiency. The fact that this is included in Component 1 further supports the contention that basic processes such as prior knowledge, self-management and guesswork were being relied upon rather than careful local reading. Further confirmation of this was found in the fact that lexical resources loaded significantly low on this component (less than 0.3 see Chapter 4 Section 4.5 and Chapter 5 Section 5.3).

There are also issues related to the validity of how the test items were constructed, particularly the use of ineffective distracters in MCQs (see Chapter 5 Section 5.2.2). This relates to a plausible explanation of guesswork as the process that was likely used in a situation where two of the three options were antonyms and the third option was not related in the context of the question. It was possible for test takers to exclude the unrelated option thus increasing their chances to 50% of answering correctly by guessing without reference to the text.

The particular test of grammatical resources relied on the test taker's ability to overcome any lexical deficit by means of drawing on syntactical parsing (see Chapter 2 Section 2.7.2.1 and Section 2.7.3.5). The fact that there was no significant difference between the group who used the intended process and those who used some other process implies that the former group were not able to effectively draw on syntactical resources (see Chapter 5 Section 2.2.3).

The evidence in the findings of Hall and Durán (2009) have shown that more proficient readers displayed the ability to overcome lexical deficit through

inferring meaning by direct semantic mapping (see Chapter 5 Section 2.2.3). It is clear from this research that most of the test takers were unable to do so effectively. It is likely that they attempted to answer the test question by reference to their L1 and were not yet able to think in English. Arabic is less opaque than English in both phonemes and graphemes (Chapter 2 Section 2.7.3.5 and Chapter 5 Section 2.3). However, proficient L2 readers are capable of overcoming this difficulty by relying on syntactical resources. For progression to academic study through English, test takers at level four should be able to draw on grammatical resources but, it is clear in this research, that most of them had not yet developed this capacity. This echoes the earlier conclusions reached by Stanovich (1980) that less proficient readers were more sensitive to contextual features. It is clear, therefore, that most of the test takers were unable to effectively draw on grammatical resources to overcome lexical challenges.

Careful global reading involves going beyond a sentence or phrase in order to pick up clues for inferring the meaning of a particular word in a sentence (see Chapter 2 Section 2.3 in Khalifa and Weir's (2009) model). In other words, ideas and details elsewhere in the text are interrelated and can be used to infer the meaning of an unfamiliar word. Although 65% of the test takers used careful reading, they were no more likely to answer the question correctly than those who used an alternative process (see Chapter 4 Section 4.6.1.4 and Chapter 5 Section 5.2.4). In fact only 54% of those who used the intended process answered the test question correctly. Alternatives to using careful global reading, very likely included guesswork to infer the meaning of 'extensively'. Of course, it is likely that some test takers may have already known the meaning of the term. However, those who used some form of guesswork may have been led astray by the three options (distractors), one of which contained the adverb 'occasionally' and the other two consisted of adverb phrases i.e. 'not very much', 'a lot'. The single word correspondence (adverb form to adverb form) may have led some to answer incorrectly.

Another possible alternative was to rely on careful local reading rather than careful global reading. Those who did so would have been expected to be significantly less successful in answering the test item. The fact that there

was no significant difference between the two groups calls into question the validity of the test item as it fails to discriminate between those who used the intended process and those who did not.

Bernhardt (1991) cited in Randall (2009) made the observation that eye movements of L2 readers tended to fixate on functional words whereas L1 readers' eye fixation tended to focus on content words. Bernhardt (1991) further suggested that L2 students with weak syntactical knowledge needed to pay much more attention to functional words, for example the adverb "extensively" in this study. More proficient readers were likely to use global reading to arrive at the likely meaning of 'extensively' if this was not already part of their vocabulary. It is likely that the less proficient readers, being more fixated on functional words, did not use the intended process of going beyond the sentence for additional information which would have helped them to infer the meaning of the unfamiliar word. Although a significant difference would have been expected between the two groups, this did not occur. In fact, the test takers who used the alternative of reading at careful local level were just as likely to answer correctly and therefore the validity of this test item is in doubt.

8.2.1.2 Component II: Expeditious reading

Expeditious search reading is a process of quickly locating required information at global level (see Chapter 2 Section 2.3). It was clear, from the high rates of agreement on the questionnaire, that most test takers had selected the intended process (see Chapter 4 Section 4.6.2 and Chapter 5 Section 5.3.1). It was also clear that those who chose the intended process had scored significantly higher on the test items than those who had chosen an alternative process. Those who had chosen an alternative process were significantly more likely to have answered incorrectly. This is strong evidence in favour of the validity of test items measuring expeditious search and there is confidence in inferences drawn about the test takers' abilities to be able to apply expeditious search effectively at academic level (see Chapter 5 Section 3.1). Support for this is found in Khalifa and Weir (2009) who pointed out that, in the case of short answer type questions, there was greater

certainty that the results were due to comprehension than to any other factor (see Chapter 2 Section 2.7.2.1 and Chapter 5 Section 3.1).

One problem with short answer type questions is their reliance on writing the answer in an item which is designed to test reading and comprehension. However, based on the evidence of the various statistical tests (see Chapter 4 Section 4.5.2.8 and Chapter 5 Section 5.3.1), this did not have an effect in the case of the test items on the Omani test. Additionally, the passage length in this particular test (and generally in other tests used in Oman) was too short and a longer passage would have been a more robust test for expeditious reading (see Chapter 5 Section 5.3.1, Chapter 2 Section 2.7.2.6)

The next feature under expeditious reading was discernment. In the Khalifa and Weir model (2009), goal setting was suggested as the means by which the reader came to a decision as to which type of reading, either careful or expeditious, would be appropriate in a given context (see Chapter 2 Section 2.3). However, in this study it was claimed that more was involved than a simple or automatic choice. In fact, choosing the appropriate type of reading for a given situation was claimed to be a high cognitive skill for which the term “discernment” was more descriptive than “target setting” (see Chapter 2 Section 2.3 and Chapter 5 Section 5.3.2). The term was introduced in Chapter 2 where, drawing on the etymology of the word, it was found to have connotations of splitting and sifting, what Purpura (1999) referred to as a self-management or executive capacity. Thus, rather than viewing it as a simple choice of target setting, it was decided to measure discernment as a variable in this study. A pleasing result of this study was that those test takers who used discernment as the intended strategy were also significantly more likely to have answered correctly (see Chapter 4 Section 4.6 and Chapter 5 Section 5.3.2). Thus, the validity of the test items measuring discernment was established, as also was the validity of the inferences drawn about test takers’ preparedness for progression to academic on the basis of these test item results. The finding that Omani test takers in this study were applying discernment appropriately was an indication that they had already begun to develop their ability to use “an organising frame” (Grabe, 2009) or “grouping relationships” (Green, 2014) for purposes of

comprehension. This is considered to be an important strategy characteristic of engaged readers (Grabe, 2009, see Chapter 2 Section 2.3 and Chapter 5 Section 5.3.2). However, it is admitted that the passage used in this test was short and that only a singular passage had been used whereas, in academic contexts, passages would be more likely to be longer and multiple (see Chapter 2 Section 2.7.2.6). Considering also the need of these test takers to develop their basic processes (based on the conclusions relating to Component 1 in Section 8.2.1.1 above), it is admitted that discernment was here being applied in a very limited set of circumstances. Nevertheless, a pleasing conclusion of this study is that many of the Omani test takers were already developing the metacognitive skill of discernment.

Finally, rubric refers to the test item instructions (task setting in Weir's (2005) framework). Although there is no direct connection between the rubrics in a test situation with that of academic reading, indirectly, one could consider that academic texts have a latent rubric in that academic readers are still reading for a purpose which is part of the task setting adopted in this study from Weir's (2005) context validation framework. Clearly, being able to use the rubric in a test task is important for developing the ability of identifying the latent rubric of an academic reading. Those who followed the instruction in the rubric were found to be significantly more likely to answer the questions correctly than those who did not follow the rubric (see Chapter 4 Section 4.6.2 and Chapter 5 Section 5.3.3). Although there was an issue of clarity relating to how detailed the answer needed to be, there was still no doubt about the validity of the test items. The lack of instruction as to whether the answer required should be short or long had implications for fairness, as some test takers were found to have given correct answers, but long ones, where long answers would add nothing substantial to what was found in the short answers.

In summary, in answering the first research question, which related to the cognitive processes by which the test takers engaged with the texts and test tasks, 34 variables were used to measure these processes (see Chapter 4 and Chapter 5). Correlation matrix results established that these variables were measuring distinct features. Whilst it is important to acknowledge that it

is difficult in a short test of 40 minutes duration to test every single cognitive feature, the research found, through factor analysis, that all the features could be categorised as either Component 1 (basic processes) or Component 2 (expeditious reading). The research found that the cognitive processes categorised under Component 1 were not being effectively applied for comprehension and were therefore underdeveloped on the part of the test takers in this study. These included scanning expeditiously, careful global and the knowledge and use of grammatical resources. However, this study did find that the cognitive processes categorised under Component 2 were being used effectively for purposes of comprehension. These included expeditious search, discernment and the use of instructions in rubrics.

8.2.2 Contextual features of LEE reading compared with those of FYA reading.

The second and third research sub-questions were designed in order to analyse test tasks and assess and compare them with an analysis of a sample of academic texts representative of the different program areas at the Colleges of Technology in Oman. The research questions were stated as follows:

2. How closely do the reading texts in LEE reflect those encountered at FYA level?

3. How closely do the reading tasks in LEE tests reflect those encountered in FYA texts?

To address both sub-questions, it was necessary to base the investigation on those contextual features (linguistic demands and task setting) identified under context validity by (Weir, 2005; Khalifa and Weir, 2009) as impacting on performance in reading tests (see Chapter 2 Section 2.6 and 2.7).

A number of different software analytical tools were used in answering these research sub-questions. These included Coh-Metrix, VocabProfiler, and WordSmith Tools (see Chapter 3 Section 3.5). Additionally, as certain aspects of texts were not directly amenable to measurement by automated

software, expert judges were asked to carry out assessments in these cases using a specially designed checklist. The items on the checklist were based on the contextual features adopted from Weir's validation framework (2005) and Khalifa and Weir's reading model (2009). The findings of these different methods were presented in Chapter 6 and further discussed in Chapter 7. The main conclusions drawn from the findings and discussion are presented in this section, taking each contextual feature in turn.

8.2.2.1 Linguistic demands

This section presents the conclusions drawn from the analysis and discussion of the linguistic demands, e. g. grammatical resources, lexical resources and writer-reader relationship. The order in which the conclusions are presented follows the order in which the findings were presented in Chapter 6 and discussed in Chapter 7.

Grammatical resources were assessed by means of 6 features namely vocabulary, grammar, readability and cohesion, and qualitatively, based on cohesion and grammar.

In terms of *vocabulary,* the test passages (LEE) were found to be representative of first year academic texts (FYA) with the exception of those variables which properly belonged to more academic technical language or discourse (see Chapter 7 Section 7.2.1). This was an important result as it provided evidence that the test passages were representative of vocabulary found at academic level with the exception of more advanced and technical vocabulary which would be encountered once academic learning had begun (see Chapter 7 Section 7.2.1).

Under *grammar,* certain aspects in the test texts were found to be representative (Chapter 7 Section 7.2.2). These included *syntactic pattern density*, *syntactic complexity and lexical density*. However, in terms of the average number of words per sentence, average number of sentences per paragraph and mean number of words before the main verb of the main clause, the test texts were found to be unrepresentative. Nevertheless, the conclusion is very similar to that relating to vocabulary in that grammar,

which is more characteristic of academic texts, was unrepresented in the test texts. The assessment by expert judges found that, on *cohesion,* test texts were representative of academic texts. Thus, the judges' assessments relating to *grammar* confirmed the earlier results found by software analysis.

Using three measures of *readability* (Chapter 6 Section 6.3.3 and Chapter 7 Section 7.2.3) test texts were found to be representative of academic texts although it is also recommended that other factors such as semantic and syntactic complexity should be considered (Weir et al., 2009).

The next contextual feature was discourse mode which refers to the way in which a text is structured in order to aid comprehension and is generally assessed by using measures of cohesion (see Chapter 7 Section 7.3 Chapter 6 Section 6.3.4). 9 different variables were used to measure cohesion and, in the case of 5 of these, no significant difference was found between LEE and FYA indicating that on these measures of cohesion LEE texts were representative of FYA texts (see Chapter 7 Section 7.3.1). It is not expected that, on measures of cohesion, FYA would greatly differ from LEE texts. It is only at more advanced academic stages that such differences would be expected (McNamara et al., 2012). However, some variables measuring cohesion were found to show significant differences. These were on measures on conceptual cohesion where LEE passages were found to be less cohesive than FYA texts although the difference was only in terms of degree of cohesion (see Chapter 7 Section 7.3.1). In general, then, LEE texts were found to resemble FYA texts on most measures of cohesions except for measures of conceptual cohesion where FYA texts were found to display a greater degree of cohesion than LEE texts. These variables were measured using automated software analysis.

A pleasing result is that the more qualitative findings, based on the perceptions of expert judges, tended to support the conclusions reached based on the automated software analysis. On two of these, *rhetorical task* and *pattern of exposition*, LEE passages were found to resemble extracts from FYA texts. In terms of rhetorical task, the passages used in both LEE and FYA were adjudged to fall into the category of *exposition*. This was not a

surprising result, as elementary college texts tend to be informative in the sense of transmitting knowledge, which would also be a characteristic of texts used in LEE (Enright et al., 2000). Similarly, for pattern of exposition, the judges tended to view both LEE and FYA texts as mainly *elaborative* in pattern although this was truer of LEE passages (Chapter 7 Section 7.3.1). However, it was noted as surprising that so few FYA texts were classified under *cause-effect* or *problem-solution* or *compare-contrast*. The point is that the LEE passages resembled the FYA texts in terms of pattern of exposition due to the fact that elementary academic textbooks still tended to be descriptive in nature. Problem-solution type or cause-effect type tend to characterise more academic texts where integration of information is required (e.g. Grabe, 2009).

On the final two variables assessed by the expert judges, *genre* and *rhetorical organisation*, significant differences were found which led to the conclusion that, in the case of these variables, LEE passages were not representative of FYA texts (see Chapter 7 Section 7.3.1 and Chapter 6 Section 6.4.1 and Section 6.4.2). The judges identified LEE passages as belonging to the *magazine or newspaper* article genre type. FYA texts tended to be classified as *textbook* genre type, an unsurprising result in that the selected texts were drawn from academic textbooks (Chapter 7 Section 7.3.1). This conclusion is similar to that of Moore et al., (2011) who found that passages used in IELTS texts failed to reflect the more pragmatic and critical genres of academic studies. In terms of rhetorical organisation, the difference between LEE and FYA texts were in terms of degree of explicitness. LEE passages were adjudged to be somewhat explicit whereas FYA texts were more explicit (see Chapter 7 Section 7.3.1). This would result in LEE passages being more difficult to comprehend than FYA passages. A number of studies concluded that comprehension was more effectively aided by problem-solution, causation or comparison than by descriptive or any other type (Khalifa and Weir, 2009). This conclusion is similar to that drawn in the case of explicitness under grammatical resources where LEE passages were adjudged to be somewhat explicit and FYA texts were found to be more explicit (see Chapter Section 7.3.1 and Section 7.7.2).

The next contextual feature, nature of information (abstractness) refers to the relative abstractness or concreteness of the information which is conveyed (see Chapter 6 Section 6.3.5, Chapter 2 Section 2.7 and Chapter 7 Section 7.4). The concreteness was measured by four variables and, in the case of three of these variables, LEE passages were found to closely resemble FYA texts. These were *measures of the concreteness of concrete words*, *the minimum concreteness within a sentences* and *mean hypernym values of verbs*. Although FYA texts contained slightly more abstract words than LEE passages, this difference was not significant and, generally speaking, approximately 75% of the words in both passages were measured as concrete (see Chapter 7 Section 7.4). It is only at a more advanced academic level beyond the first year that the language tends to become more abstract (Coady, 1979). The research questions are only concerned with the extent to which LEE texts reflect those encountered at first year academic level. On the first three measures of concreteness, no significant difference was found between LEE and FYA texts. On the remaining variable, *mean hypernym value of nouns*, a significant difference was found. FYA texts were found to have a significantly higher hypernym value of nouns than LEE passages (see Chapter 7 Section 7.4). This indicated that the academic texts contained more specific terms than the LEE passages. Although at variance with the findings of Green et al. (2010), this conclusion is internally consistent with other findings in this study, such as for grammatical resources, where the academic texts were found to be more specific and to contain more technical terms than LEE texts (Chapter 7 Section 7.2.1 and Section 7.4). Another result for the internal consistency of the current study was found in terms of cohesion, where FYA texts displayed stronger degrees of specificity than LEE texts (see Chapter 7 Section 7.3.1 and Section 7.4). However, generally in terms of abstractness/concreteness, LEE texts were found to closely resemble FYA texts, with the exception of the specificity of nouns, which was found to be higher in FYA passages. The latter conclusion actually indicates that LEE texts may tend to contain a slightly higher level of abstract terms than was found in FYA texts.

One of the linguistic demands prominent in the Khalifa and Weir (2009) model and the Weir (2005) validation framework is the illocutionary force of a passage referred to as functional resources. Here, the meaning of a passage cannot be comprehended through simply relying on the literal meaning of the words (see Chapter 2 Section 2.4, Chapter 7 Section 7.5). This use of language is not directly amenable to software analysis and was assessed in this study through the decisions of expert judges (see Chapter 6 Section 6.4.3 and Chapter 7 Section 7.5). Meaning is conveyed through a number of devices, which include *ideational*, *manipulative*, *heuristic* and *imaginative* uses of language. These are important aspects of language involving high metacognitive skills (see Chapter 2 Section 2.7.3.4). Based on the decisions of the expert judges, a significant difference was found between LEE and FYA texts. In both cases, texts were mainly identified as ideational or heuristic with a small number of FYA texts also containing manipulative and imaginative devices (see Chapter 6 Section 6.4.3 and Chapter 7 Section 7.5). LEE passages consisted mostly of ideational type functional resources whereas FYA texts were most often adjudged to be heuristic in nature. Thus, it is concluded that LEE passages did not reflect the functional resources which most often characterised FYA texts (see Chapter 7 Section 7.5).

The next feature that was examined was content knowledge, which is a linguistic feature that refers to the extent to which the comprehension of a passage relies on the background knowledge of the reader (Chapter 2 Section 2.7.3.8). In general, it is preferable that texts used for English test purposes should not be dependent on background knowledge for comprehension as this might cast doubt on the validity and fairness of the test (see Chapter 2 Section 2.7.3.8). Four variables were used in this study to identify the type of content knowledge which characterised a given passage (Chapter 6 Section 6.4.5). On three of these, *cultural, language and religion background,* no significant difference was found between LEE and FYA texts. Thus, on these three variables, it is concluded that the tasks in the LEE tests reflected the content knowledge features which would be encountered at FYA level. Academic texts tend to be less reliant on these three variables for comprehension purposes (see Chapter 7 Section 7.6). It

was, therefore, a pleasing result to find similar results for these three variables in LEE passages.

A significant difference was found for *subject specificity*. FYA texts were found to require subject specific content knowledge for comprehension whereas LEE passages tended to be extracts from newspapers or magazines which were less reliant on subject specificity for comprehension. So in summary, LEE passages reflected the FYA texts in not being over representative of cultural, language and religion background but differed from FYA by being less reliant on subject specificity.

The next feature examined was overall text purpose, which is a linguistic demand that refers to how knowledge of the purpose for which the text has been written becomes an aid to comprehension. Five distinctive types of text purposes were identified in the literature review (see Chapter 2 Section 2.1) and were used in the checklist by expert judges in (Chapter 3 and 7 Section 7.7): *referential, conative, emotive, poetic and phatic*. Almost all LEE texts were adjudged to be referential in nature (intended to inform) reflecting the very high level of referential text purpose found in FYA texts. Thus, in terms of overall text purpose, LEE texts closely reflected the referential text purpose found in FYA texts (see Chapter 7 Section 7.7). This means that the LEE in Oman was reliable indicator of students' ability to comprehend texts at academic level by means of drawing on the overall text purpose. This result is consonant with the conclusion relating to rhetorical task of an expository nature typifying both LEE and FYA passages and also, for pattern of exposition, where passages from both LEE and FYA were found to be descriptive in nature.

The next feature that was examined was writer-reader relationship. This is a linguistic demand based on the role of the text as mediating between writer and reader (see Chapter 2 Section 2.7.3.2). Two types of audience were considered in this study, namely the audience addressed and audience invoked (Ede and Lunsford, 1984, see Chapter 2 Section 2.7.3.2, Chapter 6 Section 6.4.6 and Chapter 7 Section 7.8). The recognition of this distinction in a written text is most important as an aid to comprehending the meaning of the text (Chapter 2 Section 2.2). The expert judges found no significant

difference between LEE passages and FYA texts where, in both cases, they identified that over 90% of the texts were aimed at the audience addressed (see Chapter 6 Section 6.4.6 and Chapter 7 Section 7.8). This result was what might have been expected of passages, which would be directly informative rather than rhetorical, and confirmed the earlier result related to the referential nature of texts in both LEE and FYA. Although LEE passages were representative of FYA texts in terms of audience addressed, test takers would need to be aware of the rhetorical device involved in the audience invoked –a fictional audience rather than the reader.

8.2.2.2 Task Setting

This section presents the conclusions drawn from the analysis and discussion of task setting. The task setting in Weir's (2005) validation framework and Khalifa and Weir's (2009) model include the following features:

- Response method
- Weighting
- Knowledge of criteria
- Order of items
- Channel of presentation
- Text length
- Time constraints

However, the order in which the conclusions are presented here follows the order in which the findings were presented in Chapter 6 and discussed in Chapter 7.

Despite their importance for validity, these features have not been empirically tested by the studies of Weir et al. (2009), Green et al. (2010) and Weir et al (2012) as they lay beyond the scope of those particular studies (Chapter 2 Section 2.8). In practice, it is difficult to find comparators for some of these task setting features in academic texts. However, in the current study, two of them, namely channel of presentation and text length, were tested.

Channel of presentation is a characteristic of a task whereby understanding the meaning of the task is aided by other features such as diagrams and images. It was possible to find passages in FYA texts which could be used as comparators for LEE passages.

Based on the judges' decisions, a difference was found between LEE passages and FYA texts (see Chapter 6 Section 6.5.1). However, this was a difference of degree of appropriateness with test passages being adjudged to be somewhat appropriate whereas academic texts were adjudged to be appropriate. Nevertheless, it is concluded that LEE passages were not representative of FYA tasks in terms of channel of presentation. Almost all of the test passages used in the study did not have accompanying features such as graphs, charts or even headings. There was just a single case where a passage contained a table. Almost all of the academic texts, in contrast, included additional features such as charts and images. Despite what is claimed in the literature, that comprehension is greatly aided by information being presented in more than one channel (see Chapter 2 Section 2.5), LEE passages were found to rely on a single channel, that of written text (see Chapter 7 Section 7.91).

The second task setting is text length. Again a difference was found based on the expert judges' ratings between LEE and FYA texts regarding the appropriateness of the passage in terms of length. FYA texts were adjudged to be appropriate whereas LEE passages were somewhat appropriate (see Chapter 6 Section 6.5.2). However, this difference was found to be non-significant. Accordingly, it is concluded that, in terms of text length, LEE passages were representative of FYA texts. However, there is an issue related to the number of passages used since, in the Omani situation, only a single passage was used making it impossible to test comprehension through organising frames. So, even though in terms of text length LEE passages were found to be representative of FYA texts, there is a need for more than a single text for testing comprehension through organising frames (Grabe, 2009).

In summary, generally LEE texts and tasks were found to be representative of FYA texts. Comparisons were made based on measures of vocabulary, grammar, readability, cohesiveness, nature of information, linguistic demands and task setting. Where significant differences were found, it was usually the result of the test texts being more general in nature in comparison to the FYA texts, which tended to be more subject-specific. For example, on vocabulary measures, where differences were found, they were on those variables that measured the third 1,000 words and sub-technical vocabulary. FYA texts were found to contain a higher number of words from the third 1,000 words or sub-technical vocabulary than LEE passages. However, in terms of grammar, LEE passages were found to be significantly higher in terms of word per sentence and sentence per paragraph than FYA texts. What this means is that LEE texts tended to have longer sentences and paragraphs than FYA texts which were often more terse. However, this indicates that, in many cases, LEE passages were more challenging in terms of density of logical operators than FYA texts. On measures of readability, LEE passages were generally representative of FYA texts with one exception where a difference was found on the R2 measure (Flesch-Kincaid Grade level (0-16)). However, it has been shown that this difference has little meaning due to the fact that both LEE and FYA texts fell below a threshold which was set in the US and which might not be appropriate in other contexts. On measures of cohesiveness, the only difference which was found was in terms of degree of cohesiveness where FYA texts were more conceptually cohesive and contained more specialist vocabulary than LEE passages. Similarly, in terms of nature of information, FYA texts displayed greater specificity than LEE passages.

The difference, in terms of subject specificity, is perhaps best exemplified in terms of genre where the LEE passages tended to be extracts from magazines and the FYA texts were more of academic research type. This, at least partially, explains why, on functional resources, LEE passages were found to be more ideational and descriptive in nature whereas FYA texts tended to be more heuristic in nature. However, on the other linguistic

demands such as overall text purpose and writer-reader relationship, LEE passages were found to be representative of FYA texts.

One significant difference between LEE and FYA texts was in terms of the task setting of channel of presentation where LEE tasks relied on the single channel of written texts in contrast to FYA texts which used many channels such as charts and graphs to support the text. The other task setting examined here was text length where LEE passages were adjudged to be representative of FYA texts.

Thus, it is clear that LEE texts and tasks generally reflected those found in FYA although the LEE texts and tasks tended to be more general in nature and did not reflect the technical and subject specificity characteristic of FYA texts.

8.3 Implications

8.3.1 Implications for test theory

Over recent decades a number of models of language testing have paid attention to how reading ability could be assessed. The earlier frameworks provided by Bachman (1990) and Bachman and Palmer (1996), although quite comprehensive in scope, were critiqued for their neglect of the underlying cognitive processes involved and for their lack of empirical foundation (O'Sullivan and Weir, 2011). One model which drew on the earlier work of Bachman but which was empirically based was the socio-cognitive validation framework of Weir (2005) (Weir et al., 2009; Green et al., 2010). This framework was adopted due to its usefulness in the context of this research, the main objective of which was to find a set of clearly defined variables by which it would be possible to ensure the context validity of reading tests for academic purposes. Such a set of clearly defined variables was provided by Weir's framework (see Chapter 2 Section 2.5).

However, in the earlier work of Urquhart and Weir (1998), an important distinction was made between two types of reading namely careful and expeditious. Comprehension based on either careful or expeditious reading

could occur at either local or global level in academic reading. An important development of Urquhart and Weir's model (1998) and the validation framework of Weir (2005) was the model of reading provided by Khalifa and Weir (2009) on which the current research was based (see Chapter 2 Section 2.3).

One of the aims of this research was to identify the cognitive processes involved in reading and comprehension in L2 (research sub-question 1). In adopting the Khalifa and Weir's model it was therefore of importance in the context of this study to establish the robustness of the model and that it was measuring distinct cognitive features. This was necessary in order to validate the model.

The need for validating existing models has been commented on by a number of authors (e.g. Fulcher, 1998; Alderson and Kremmel, 2013). With particular reference to the testing of reading, Fulcher (1998) made a strong plea for models to be empirically validated, arguing that this was necessary to avoid construct underrepresentation which he perceived as presenting a major threat to validity. In this context, Alderson and Kremmel (2013) emphasized the difficulty of devising a set of construct components embracing all of the cognitive processes which can be involved in reading and comprehension for L2. Furthermore, they stressed the importance of such a set of constructs for understanding and diagnosing the strengths and weaknesses of test takers in L2 reading and comprehension.

The findings of the correlation tests revealed that each of the cognitive features identified in Khalifa and Weir's (2009) model was indeed distinct from the others as no pair of constructs was found to be strongly correlated (see Chapter 4 Section 4.3). Factor analysis led to a two component scheme based on basic processes and expeditious reading rather than on the careful and expeditious reading of the Khalifa and Weir (2009) model. Moreover, factor analysis further supported the results of the correlation matrix regarding the distinct nature of the different features used in this study for examining the cognitive processes. This confirmed the validity of the model

for examining context validity and provided an appropriate framework for addressing the first research question (Chapter 4 Section 4.4).

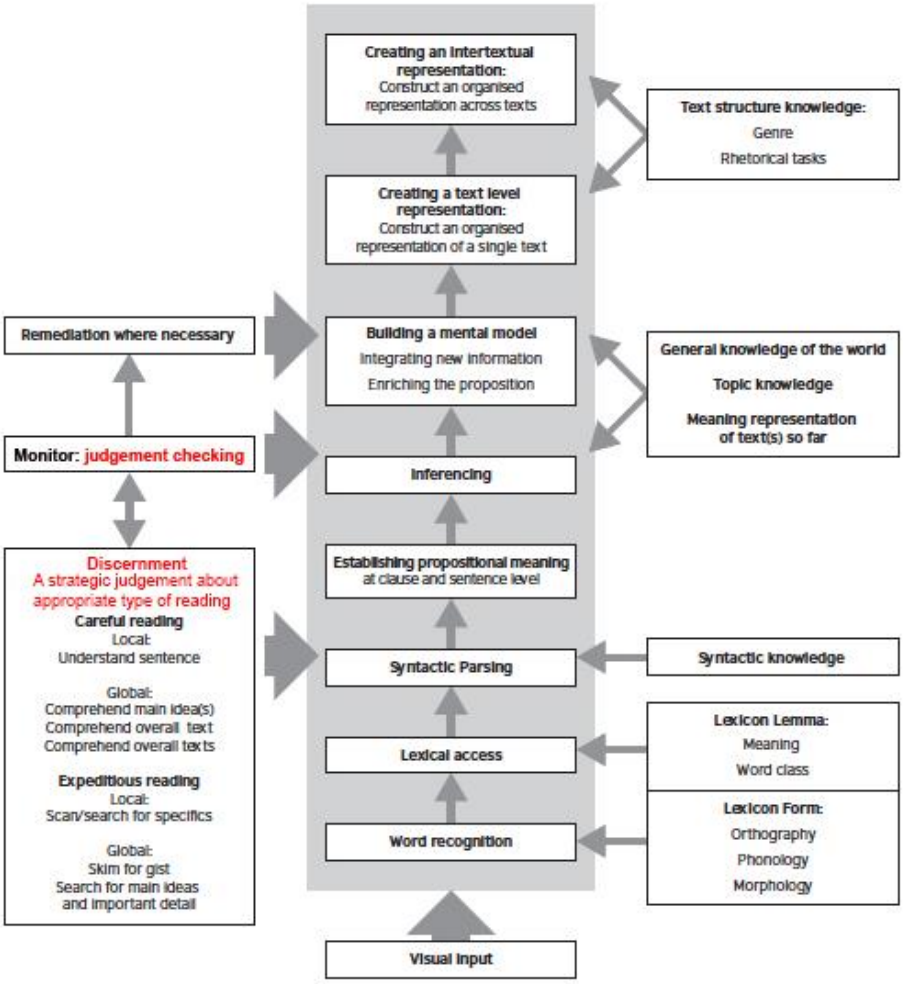
The implication of validating Khalifa and Weir's (2009) model is that the various stakeholders involved in the test can have great confidence regarding inferences made on the basis of the cognitive processes through which test takers had engaged with the various tasks in the test. Thus, the context validity of the model has been established in a L2 context. There is also confidence that similar findings would result from research into L2 reading in other contexts.

One of the strengths of the Khalifa and Weir's (2009) model is that it includes modes of reading, careful and expeditious. However, it has been pointed out that, for academic reading, much more attention has been given to careful reading to the neglect of expeditious reading, despite its importance for academic purposes and its inclusion in the EAP descriptors of the CEFR (Weir, 2013). In the current study, it was demonstrated that expeditious reading was not only a distinctive feature, but was actually one of the two components upon which the different variables loaded. Consequently, there is a need for expeditious reading to be given equal attention in research to that which has been paid to careful reading.

Another implication for test theory was the introduction of a new conceptualisation of the "goal setter" function of the Khalifa and Weir (2009) model. The model suggests that goal setting involves selecting the appropriate type of reading, careful or expeditious, in a given context (see Chapter 2 Section 2.3). The findings and conclusions of this study suggested that more is involved than "goal setting" as a simple choice between careful and expeditious reading (see Chapter 5 Section 5.3.2) as was argued in the literature review (see Chapter 2 Section 2.3). The term "discernment" with its connotations of "sifting" and "splitting" is, therefore, proposed as a concept, which implies a strategic judgment and not simply an automatic or reflex action. It implies a considered judgment involving the creation of an organising frame for comprehension (Grabe, 2009) or grouping relationships (Green, 2014). The main thrust of Khalifa and Weir's (2009) model was to

show how reading and comprehension processes, widely acknowledged to be complex activities, involved the use of various metacognitive skills, which involved the decision to read either carefully or expeditiously. That choice was simply nominated as goal setter but, as has been argued in this research, self-management, splitting and sifting, and integration are elements involved in the decision to read either carefully or expeditiously for comprehension. The proposed modification to Khalifa and Weir's (2009) model is presented in Figure 8.1:

Figure 8.1 A proposed modification to Khalifa and Weir's (2009) model



In this adapted model "goal setter" has been replaced by "discernment" which is *a strategic judgement* about the appropriate type of reading required and implies the processes of *self-management*, *splitting and sifting* and *integration*. The monitoring process, rather than simply being "goal

checking", is replaced by "judgement checking". With this adapted model in mind, underlying theories for test design ought to consider these elements of discernment (i.e. self-management, splitting and sifting and integration).

The fourth implication, from a theoretical perspective, is the design of a questionnaire to capture the cognitive processes of test takers in completing the various test tasks. The initial concept of devising a Verbal Protocol Analysis (VPA) questionnaire as an instrument arose from the study of Akmar Saidatul Zainal Abidin presented in Weir (see Chapter 3 Section 3.4.1). In that study, the author devised a questionnaire with the aim of capturing the cognitive processes involved in *speaking*. The main thrust of that instrument was to investigate the context validity features through which the test takers engaged with the various test tasks. The general idea of using a similar approach in the current research was considered although it was immediately evident that a completely different set of variables would need to be devised to measure the context validity features of reading.

The fundamental elements of the design of the instrument were as follows:

- Context validity features (Weir, 2005) and Khalifa and Weir's model (2009)
- VPA guidance from Green (1998)
- Questionnaire construction protocols from various literature
- Mapping to a sample reading test

Devising the instrument involved a rigorous process whereby 15 distinct contextual features were identified from Weir's validation framework (see Chapter 2 Section 2.7) and, taking into consideration the different modes and types of reading (careful/expeditious and global/local) according to Khalifa and Weir's (2009) model, resulted in a set of 45 variables (see Chapter 3 Section 3.4.1). These were reduced to 34 variables once allowances were made following questionnaire construction procedures and guidance (see Chapter 3 Section 3.4.1 and Chapter 4 Section 4.2). Further refinements were made to the instrument (See Chapter 3 Section 3.6.1) after a pilot study had been conducted. Following correlation and factor analysis there is

confidence in the validity of this instrument for use and adaptation for future research into the context validity of reading.

In respect of measures of readability (R1, R2 and R3) the current research confirms the findings of Khalifa and Weir (2009) and Weir et al. (2006) and Weir et al., (2009) and Green et al. (2010) where the issue was not whether the test passages were representative of the academic texts but rather that both fell below thresholds set on these measures in a US context but which might not be appropriate in other contexts (see Chapter 7 Section 7.2.3). Applying these measures uncritically leads to some anomalous conclusions such as a significant difference being found on R2 but no significance found on R3 which actually represents a more advanced measure than either R1 or R2 (see Chapter 7 Section 7.2.3). Since no significant difference was found on R3, it follows that little meaning can be attached to any differences found on R1 or R2 since both LEE passages and FYA texts, measured on either variable, fell below the various thresholds anyway. Despite the acceptance of these measures by many scholars, Masi (2002) critiqued them for their inadequacy for revealing text complexity and difficulty (see Chapter 2 Section 2.7.3.5 and Chapter 7 Section 7.2.3). Weir et al., (2009) have argued that these measures should not be considered alone without taking into account semantic and syntactic complexity. It is, therefore, recommended that R1, R2 and R3 as measures of readability should not be seen in isolation from other measures such as grammatical resources, vocabulary and grammar, each of which was examined both quantitatively and qualitatively in this research.

In summary, this research has highlighted five main implications for test theory. The first resulted from the need for the empirical validation of models of reading and comprehension as commented on by many scholars. This research has provided such validation for the Khalifa and Weir (2009) model in a L2 context. This research has demonstrated that the cognitive parameters identified in the model were distinct. Consequently, test stakeholders can have increased confidence in inferences made on the basis of this model relevant to the test takers' preparedness for academic study through the medium of English.

The second implication of this research is to emphasise the need for more attention to be paid to expeditious reading due to its neglect in research. However, this research has highlighted the importance of expeditious reading in an academic context and that it, therefore, merits more attention in research.

The third implication is the need for research into discernment, which is proposed in this study, as a more appropriate descriptor for the strategic judgement the reader must make between careful and expeditious reading. This represents a modification to the Khalifa and Weir (2009) model where "goal setter" is replaced by "discernment" as a higher metacognitive strategy. Consequently, a modification is required to the monitoring process, which becomes "judgment checking" rather than "goal checking". The different elements of discernment have been identified in this research (i.e. self-management, splitting and sifting and integration) and these merit consideration in future research.

A further implication is that the questionnaire, which was devised in this research, can be presented as an exemplar for use or adaptation by other researchers investigating context validity of reading or other skills. This instrument was devised as the result of a rigorous process following best advice on questionnaire construction. It was further refined following piloting and its validity was established by means of correlation tests and factor analysis. Thus, there is confidence in its validity for use in other related research contexts.

Finally, a contribution to test theory is made in the context of measurements of readability. It was found that the results of certain measures of readability such as R1 *Flesch Reading ease 0-100*, R2 *Flesch Kincaid Grade level 0-16*, and R3 *Coh-Metrix L2 readability* should not be taken uncritically or in isolation from other measures such as grammatical resources, vocabulary and grammar. Whilst the results of those measures can be considered to be broad indicators, this research has found that the results based on other measures provided by the automated software needed to be taken into account. Additionally, the picture of readability can be further enhanced by

taking into account the more qualitative judgements made by experts, which provided complementarity to the results obtained through the more quantitative measures.

8.3.2 Implications for test design

The findings, analysis and discussions involved in this research have a number of implications for test design. The first of these concerns the testing of expeditious scanning. It was found that there was a need for a strictly time bound task to avoid test takers fixating on single words. Such a time restriction would compel the test takers to more quickly locate required information by comprehension of a group of words rather than fixating on a single word.

In relation to the testing of expeditious scanning, Khalifa and Weir (2009) have identified the need for randomisation of test items as a requirement for validity. By presenting the test items in a logical structure and verbatim with the text itself, as was the case in test tasks in this study, the test designers have defeated the object of the exercise by allowing test takers to answer the questions by processes other than scanning (see Chapter 5 Section 5.2.1). In test design, therefore, it is recommended that items testing expeditious scanning should be presented in a randomised order and, additionally, that simple binary choices should be avoided. The validity of test tasks for expeditious scanning can be greatly improved by the provision of more than two choices in MCQ and by randomising the order of the test items.

Another issue relating to the use of MCQs is the inclusion of ineffective distractors (see Chapter 5 Section 5.2.2). It is recommended that, in constructing MCQ type questions, care should be taken to avoid having ineffective distractors, which might increase the test takers opportunities to answer correctly by guesswork without reference to the text. Equally important for validity, is the avoidance of a distractor that leads the test taker to answer incorrectly. An example in this research was based on adverb-to-adverb correspondence, which was discussed in (Chapter 5 Section 5.2.4), where the test taker might have been unfairly misled.

However, it is recommended that short answer type questions are preferable to MCQ type, especially for testing expeditious reading, as there is greater certainty that the result of such questions is due to comprehension rather than to any other factor (see Chapter 2 Section 2.7.2.1 and Chapter 5 Section 5.2). It is therefore recommended to test designers that more short answer type questions should be included and that there should be less reliance on MCQs.

It is also clear that the test item purporting to measure grammatical resources lacks validity (see Chapter 5 Section 5.2.1 Section 5.2.3). For example, the L1-L2 transfer involved in the Arabic term for 'و' is likely to have aided comprehension for those who used an alternative process and this point needs to be considered for the construction of valid test items.

Although there were no issues of validity relating to the testing of expeditious search, where the test tasks were found to properly discriminate between those who used expeditious search and those who did not, it is nevertheless recommended that there should be several passages rather than the single passage typically used in the Omani test. Table 2.5 in Chapter 2 Section 2.7.2.6 adopted from Khalifa and Weir (2009) displays a number of alternatives used by different international language tests. What is clear from that table is that the typical test used in Oman falls below the minimum number of passages used internationally. It is, therefore, recommended that tests designers consider a redesign of the passages used in line with the overall test purpose. Obviously, this also has implications for time allocation. However, it is admitted that there are no issues of validity with the current practice in this respect.

However, there is another reason for recommending more than one passage and this is that it offers more opportunities for testing discernment. Although there were no concerns about the validity of the test tasks testing discernment in this study, in academic contexts, passages are more likely to be longer and multiple (see Chapter 2 Section 2.7.2.6). In view of what has been said about discernment involving "an organising frame" (Grabe, 2009) or "grouping relationship" (Green, 2014), the inclusion of a number of

passages in a test would provide greater scope for measuring discernment in situations that more closely resembled academic contexts.

Another issue relates to under-representation of technical vocabulary and vocabulary above the 3,000 word threshold. Coh-Metrix analysis showed that test passages were under-representative of vocabulary in these categories (see Chapter 7 Section 7.2.1). The recommendation here is that test designers should be aware of the need for the inclusion of more vocabulary from these categories to enable valid inferences to be made about students' abilities to cope with the more technical vocabulary typically found in academic reading. However, the role of grammatical resources in overcoming lexical deficit must also be considered. In this respect, the findings of this study showed that LEE passages were unrepresentative of academic texts in that LEE passages tended to have longer sentences and a greater number of words per sentence. However, sight must not be lost of the scope of the academic studies in this research which were mainly technological in nature and would therefore have tended to consist of more terse and concise sentences (see Chapter 7 Section 7.2.2). Tempting though it might be to suggest that test passages should be more terse and concise to reflect the academic type texts, there is ample scholarly advice against doing so (Khalifa 1997; Alderson, 2000 and also see Chapter 2 Section 2.7.3.8). For this reason, many international tests such as IELTS changed the passages used for test purposes from subject specific to more general passages and this is also the practice in Omani tests. In doing so, they have become better tests due to being less reliant on subject specificity for comprehension. This fact was confirmed in the findings of the expert judges who concluded that the FYA texts were more dependent on subject specific knowledge than the LEE passages (see Chapter 7 Section 7.6). This is confirmed by similar findings in Weir et al (2009) and Green et al. (2010).

Nuttall (1996) and Fulcher (1999) have shown that the evidence does not support the case for greater subject specificity in tests and Weir (2005) actually found that the effect of content knowledge on comprehension was minimal. However, Alderson (2000) and Weir et al. (2009) found that topic familiarity was a significant predictor of difficulty in comprehension.

Nevertheless, it is not recommended that test passages should be more subject specific even though there is a need for building up vocabulary and syntax as is recommended for curriculum design and programme managers (see Section 8.3.3). Instead, tests should be constructed to provide a variety of texts drawn from different academic areas (Enright et al., 2000). This is supported by Cushing-Weigle (2000) who, in commenting on the MELAB proficiency test, asserted that a wide variety of test texts reflecting different subject areas would minimise the risk of some test takers being either assisted or impeded by unequal content knowledge (see Chapter 2 Section 2.7.3.8).

In relation to functional resources, it has already been noted that FYA texts consisted of significantly more specific and more abstract words than LEE passages. This is explained by the fact that students are still at foundation level in the various sciences and, consequently, passages used for their tests tended to be more general and descriptive in nature making more use of the functional resource of ideational devices (see Chapter 7 Section 7.5). As students progress to more academic levels, underlying principles are introduced and these require definitions and committing formulae to memory through the use of heuristic devices. This is in concordance with the results found earlier for genre where expert judges assessed LEE passages as mostly magazine or newspaper type whereas FYA texts tended to be mostly research type genre (see Chapter 7 Section 7.5). Thus, in terms of functional resources, LEE texts tended to be more ideational and descriptive and did not reflect the more heuristic nature of functional resources found at FYA. The implication is that this reflects the reality of the difference between foundation and FYA levels. It would not be feasible to have greater representation of heuristic functional resources in LEE passages. Also, foundation students are not yet streamed into their different disciplines. Consequently, passages used still need to be general and descriptive.

In terms of writer-reader relationship, LEE passages closely resembled FYA texts, as over 90% of passages in both tended to be audience-addressed type. However, it is recommended that test designers should incorporate some passages based on audience invoked in order to test the test takers'

ability to use this device as an aid to comprehension. This presupposes that test designers have an awareness of the audience addressed which in this case is the foundation level student and that they should have some awareness of what can be assumed to be part of the student's life experience, knowledge and other characteristics, so that what is written in the text engages with all of these characteristics but without unduly going beyond their range of experience (see Chapter 2 Section 2.7.3.2 and Chapter 7 Section 7.8).

Two further recommendations resulted from the inclusion of two task setting features in the current study. Weir at al. (2009) have drawn attention to the need for task setting features to be empirically tested. Although, in practice, this is difficult to achieve due to the difficulty of finding comparators, two task setting features were included in this study namely channel of presentation and text length (see Chapter 6 Section 6.5). In relation to the latter the need for a longer passage and multiple texts has already been recommended. With respect to channel of presentation, despite the importance attached to the use of more than one channel to aid comprehension (see Chapter 2 Section 2.7.2.5, Chapter 6 Section 6.5.1 and Chapter 7 Section 7.9.1), LEE passages are overly reliant on the use of a single channel namely written text. LEE passages were found to be unrepresentative of academic texts, which included features such as graphs, charts and headings. It is recommended that test designers should include texts supported by graphs, charts or images as aids to comprehension, as is common practice in many international tests e.g. CAEL (see Chapter 2 Section 2.7.2.5 and Chapter 7 Section 7.9.1).

In summary, it is recommended that test designers should firstly include a time-bound task in order to validly test comprehension based on expeditious scanning. Secondly, there is a need for test items to be presented in a randomised order as this offers greater assurance that expeditious scanning is being tested. Additionally, there is greater assurance of validity by the avoidance of simple binary choices in MCQs and by providing a wider range of distractors. With particular reference to distractors, the conclusions of this research imply that care should be taken to avoid ineffective distractors,

which may either aid the test takers in answering correctly or mislead them towards an incorrect answer. In either case, the validity of the test items would be in doubt. Actually, this research has shown that short answer type questions are preferable for testing expeditious reading and therefore there should be a greater balance in favour of short answer type questions and less reliance on MCQs.

This research also alerts test designers to the problem of L1-L2 transfer aiding comprehension and thereby affecting the validity of items testing grammatical resources. This problem needs to be considered in test design.

Although in the test under examination in this research there were no issues of validity related to reliance on a single text, it is, nevertheless, recommended that test designers should consider the inclusion of a number of passages which would bring the Omani test in line with most international tests such as CAEL and IELTS. Multiple texts would allow for a more robust test of discernment, as test takers would have greater opportunities for organising frames for comprehension. Additionally, having a number of texts drawn from different academic areas, would minimise any possible bias that might arise from topic familiarity.

In this research, test passages were found to be under-representative of the technical vocabulary above the 3,000-word threshold found in academic texts. It is tempting to recommend a greater representation of technical vocabulary in test passages. However, there is a need for caution in doing so, in order to avoid the test passages becoming unduly subject specific.

It is also recommended that test passages should be more representative of academic texts by the inclusion of more than one channel of presentation. The test passages under examination in this research were completely reliant on written text. There is a need for the inclusion for graphs, charts, tables and pictures as these characterise academic texts.

All of the foregoing recommendations relating to test design draws attention to the need for training for test designers/writers to become aware of these problems and their implications for validity.

8.3.3 Implications for curriculum designers and programme managers

In Section 8.2.1, conclusions were reached in answering the first research sub-question, which related to the cognitive processes through which test takers engaged with the test tasks. These processes were categorised under two main components, namely basic processes and expeditious reading. The features which loaded on Component 1 (basic processes), did not yield significant results when tested against actual test scores indicating that these skills were underdeveloped in the case of the Omani students. Accordingly, this implies that cognitive processes such as scanning expeditiously, careful reading global and the knowledge and use of grammatical resources need to receive more attention in teaching and learning on foundation programmes.

The cognitive processes in Component 2, for example, expeditious search reading, the use of discernment and the use of instructions in rubric all proved to be significant when tested against actual test scores which implied that, where test takers applied these processes, they also tended to score significantly higher on the relevant test tasks. Therefore, with reference to inferences made on the basis of these cognitive processes, there is confidence about their validity for decision-making regarding the test takers' abilities to progress to academic level.

It was found that scanning expeditiously did not load on the expeditious reading component but rather on the basic processes component. This was likely due to its underdeveloped nature in the case of Omani test takers. It is therefore recommended, that in the pedagogy of English as L2, more attention needs to be given to exercises where the meaning can be comprehended from a group of words rather than fixating on the meaning of a single, usually abstract, word. In order to achieve this, exercises that are strictly time constrained, should be devised to encourage students to develop this skill. Alderson (2000) saw the importance of this for the real word situation or for academic situations where time was limited.

However, where careful reading at local level is required, the conclusion was that the test takers' inability to apply this process was related to their lexical and syntactic deficits. A solution based on teaching subjects at secondary

level through the medium of L2 is probably not a feasible one in the Omani context as such a change to practice lies beyond the remit of the Colleges of Technology. Instead, strategies in curriculum design and teaching are necessary to improve students' lexical and syntactic knowledge. In the discussion chapter (Section 7.2.1), consideration was given to the usefulness of word lists where Khalifa and Weir (2009) suggested raising the threshold beyond the 2,000 most frequent words which represented 95% of vocabulary recognition to the 5,000 most frequent words level which would represent 98% coverage and that this was more appropriate for students advancing to academic level. This is supported by the findings of the current research where academic texts were found to be significantly more representative of technical vocabulary than LEE passages (see Chapter 7 Section 7.2.1). Furthermore, it was a finding of this research that students were not able to effectively draw on grammatical resources which could have helped them to overcome their lexical challenges (see Chapter 5 Section 5.2.3). In fact, Shiotsu and Weir (2007) claimed that syntactical knowledge was even more important than lexical knowledge for comprehension.

Accordingly, it is a task that is recommended for curriculum designers and programme managers to devise effective strategies for vocabulary and syntax building across each level of the Foundation Programme. A helpful approach for syntax building is the incremental scheme discussed in Chapter 2 Section 2.7.3.5 which commences with simple sentences and progresses to multiple sentences (compound and complex) as presented in Quirk et al. (1985), Crystal (1996), Huddleston and Pullum (2005) and Carter and McCarthy (2006). On the Foundation Programme, at lower levels, the focus can be on simple sentences based on a single independent clause. For higher levels, the curriculum can extend the students' proficiency to include multiple sentences, both compound and complex.

It has already been concluded that LEE passages in Oman were indicators of the students' abilities to comprehend texts at academic level by being able to correctly identify the overall text purpose as an aid to comprehension. However, LEE passages resembled the FYA texts in terms of *referential* type

purpose (see Chapter 7 Section 7.7). Nevertheless, it is recommended that, in teaching and learning, other types of text purposes such as conative, emotive and phatic should be encountered by the students due to their prominence in various academic contexts. For example, conative and emotive purposes are often found in marketing strategies in business studies and conative purposes may be encountered in legal aspects of certain courses. Even more important, students need to be able to distinguish between factual texts based on evidence and texts based on opinion. Additionally, it has already been shown in Chapter 2 Section 2.7.3.1 and Chapter 7 Section 7.7 that the more advanced tests such as Cambridge Advanced English (CAE) and Certificate of Proficiency in English (CPE) incorporated a wider variety of test purposes than referential purposes. These additional test purposes included poetic, emotive and conative (Khalifa and Weir, 2009). In fact, the move towards more written communication involving the use of chat rooms and blogs, implies that contributions might not always be referential in nature. This is suggested by Khalifa and Weir (2009), who have emphasized the growing importance of more phatic and emotive purposes of texts in the social media. The implication of this is that, even though the LEE passages resembled the referential purpose type of FYA texts, recognising other types of purposes are becoming increasingly more important for the more collaborative ways in which the students now learn at academic level.

In summary, the findings of this research have led to a number of recommendations for curriculum designers and program managers. The first is that, in teaching and learning, more attention should be given to the cognitive processes of scanning expeditiously, careful reading at global level and the knowledge and use of grammatical resources for comprehension. Secondly, in the teaching of English as L2, greater attention needs to be given to exercises where the meaning can be comprehended from a group of words rather than fixating on the meaning of a single word, which is often abstract in nature. For this to be effective pedagogically, the exercises need to be strictly time constrained. However, this does not detract from the need for improving lexical and syntactic knowledge at all levels of the foundation

program. Finally, the importance of identifying the overall purpose of texts is already recognised within the foundation program. However, this research found that this was mainly confined to the referential type of purpose. Students at academic level need to be able to recognise other purposes of texts such as conative, emotive and phatic as important aids to comprehension. Additionally, in view of the growing importance of written communication through the social media, it is important that students understand the use of phatic and emotive purposes in texts.

8.4 Limitations and pointers for further research

There are a number of limitations inherent in this research and these should be borne in mind when considering its conclusions and implications. These are now presented and briefly discussed.

The Khalifa and Weir (2009) model consists of three types of context validity parameters, namely task setting, linguistic demands and task administration. The first two were addressed in this research but not the third type (see Chapter 2 Section 2.5). Task administration is important for assuring context validity and merits special attention beyond the scope of this study. It focuses mostly on the practical aspects of exams administration and includes such variables as examination timetable and location, physical conditions of the exam venue, security and exam materials. Although these are well established as regulations governing examinations generally, there is a need for a study of task administration alongside task setting and linguistic demands in the context of testing of L2 reading and comprehension.

The second limitation relates to the difficulty of assessing certain task setting features. This difficulty lay in the problem of finding suitable comparators in FYA texts for particular LEE test tasks, for example, time constraint. In reality, at academic level, one could spend a whole night on understanding a passage so that the time constraint in a test context has no real life comparator in academic reading. The limitation is not only that of a lack of a real life comparator but also that measuring it as a variable in this study relied on the subjective judgment of experts. However, the judges were

expert and experienced in teaching, learning and testing of L2 (see Chapter 3 Section 3.5.2). Furthermore, in comparing each set of LEE tasks and FYA texts, the opinions of three judges were used so that the problem of subjectivity was likely to be overcome by inter-subjectivity evidenced by the generally high rates of agreement between the judges, indicating high levels of inter-rater reliability (see Chapter 6 Section 6.4). Similar limitations of comparison apply to weighting parameter in this study and other task setting features such as response method, knowledge of criteria, order of items and time constraints were not included in this study for that reason. With regard to the task setting features that were included, it is admitted that the test tasks relating to them were limited in resembling real life academic situations but nevertheless the findings and conclusions were reliable due to the high levels of inter-rater reliability (see Chapter 6 Section 6.4). However, there is a need for research into methods of measuring task setting features for purposes of comparison.

A third limitation is that the natural experiment was conducted by means of a simulated rather than a real test. Due to the high stakes nature of the real test and the fact that the reading test was part of a more comprehensive test, which included listening and writing, a real time observation was considered to be intrusive and might have affected the validity and fairness of the test itself. The reasons for conducting VPA using a simulated test have been discussed in Chapter 3 Section 3.4 where a number of alternatives were considered to be impractical for the current research. The purpose of this experiment was to capture the cognitive processes by which the test takers engaged with the test tasks and there is confidence in the results obtained due to the results of correlation and factor analysis (see Chapter 4 Section 4.3 and Section 4.4). Related to this limitation is also the fact that only a single version of a test and questionnaire was used but the practicality of using numerous scripts would not have been feasible within the scope of this study. However, as has been earlier recommended, a validated and piloted questionnaire has been devised for this study and it is recommended that its adaptation in other contexts will further confirm its usefulness.

A fourth limitation lies in the sample size used in the natural experiment which was designed to answer the first research question related the cognitive processes through which students engaged with the test tasks. The sample size of 202 test takers is relatively small in comparison to a student population of over 11,000. However, as this was a natural experiment (see Chapter 3 Section 3.4 and Chapter 4), it relied on the students' willingness to undertake a simulated test, which had to be set up and cleared through various levels of administration and also involved marking the exam papers. Within these practical constraints, a sample size of 202 was, considered to be feasible given that this was one method among others. However, despite the small sample size, there is confidence in relation to internal validity based on correlations and factor analysis.

8.5 Summary

In addressing the first research question, the model of Khalifa and Weir (2009) was validated in a L2 context, and the various features of reading as a multi-componential skill were confirmed as distinct parameters for context validity. In this context, factor analysis revealed a two-component model based on basic reading processes and expeditious reading as categorising the various cognitive features by which the test takers engaged with the test passages. It was found that the first component, consisting of basic processes such as careful reading at local level, the knowledge and use of grammatical resources and expeditious scanning, were reading skills which were still largely underdeveloped in the case of these L2 test takers who were learning English for academic purposes in a context where English was not the first language. However, the study did reveal that skills of expeditious reading such as expeditious search, discernment and the use of instructions in rubrics were well developed and aided comprehension in the case of most of the test takers. In validating the Khalifa and Weir (2009) model, a modification was proposed in relation to the goal-setting element. It was suggested that discernment would be a more appropriate term for conveying the more sophisticated cognitive skills involved in making a strategic judgement between careful and expeditious reading.

The second and third research sub-questions involved an assessment of how closely the test passages resembled academic texts. This was achieved through the use of software such as Coh-Metrix, VocabProfiler, and WordSmith and also the decisions of expert judges, for those features that were not directly amenable to measurement using the quantitative software. In general, the test passages were found to be representative of academic texts and, although the nature of the test passages did not reflect the subject specificity and abstractness characteristic of academic texts, a strong argument supported by many authors in the field of L2 testing was presented in favour of retaining the more general nature of test passages as valid for testing reading skills.

A number of recommendations were made both for theory and practice. The need for further research into test administration features and also for the development of techniques for investigating task setting features were emphasised. Implications for test design included the need for less reliance on MCQs and more use of short answer type questions for increased context validity. For teaching and learning, it was recommended that greater attention be given to syntax and grammar to overcome the difficulty posed by lexical deficit in comprehension.

It is hoped that the recommendations made in this study will contribute to the scholarly discourse relevant to testing reading in English for Academic Purposes (EAP), test validation studies and the extension of such studies to the testing of other English language skills. A desirable outcome of this research would be the instigation of further evidence-based studies, the findings of which would lead to an enhancement of L2 teaching and learning and informed decision-making about students' performance. There is great confidence that the insights gained through this research will lead onto improvements in language teaching, learning and testing, and that these, in turn, will enable students to communicate more effectively in English in real-life contexts and also contribute to the greater good of society at large.

References

Adams, M. J. (1990) *Beginning to read: thinking and learning about print*. Cambridge, MA: MIT Press.

Adams, M. J. (2004) Modeling the connections between word recognition and reading. In: Ruddell, R. B. and Unrau, N. J. (eds.) *Theoretical models and processes of reading*. 5th edn. Newark, DE: International Reading Association. pp.1219-1243.

Akbari, R. (2012) Validity in Language testing. In: Coombe, C., Davidson, P., O'Sullivan, B. and Stoynoff, S. (eds.) *The Cambridge guide to second language assessment*. Cambridge: Cambridge University Press. pp.30–35.

Alderson, J. C. (2000) *Assessing reading*. Cambridge: Cambridge University Press.

Alderson, J. C. (2005) *Diagnosing foreign language proficiency: The interface between assessment and learning*. London: Continuum International Publishing Group Ltd.

Alderson, J. C. (2009) Test review: Test of English as a foreign language: Internet-based Test (TOEFL iBT). *Language Testing*, 26(4), pp. 621-631.

Alderson, J. C. and Urquhart, A. H. (eds.) (1984) *Reading in a foreign language*. London; New York: Longman.

Alderson, J. C., Clapham, C. and Wall, D. (1995) *Language test construction and evaluation*. Cambridge: Cambridge University Press.

Alderson, J. C., Figueras, N., Kuijper, H., Nold, G., Takala, S. and Tardieu, C. (2004) *The development of specifications for item Development and*

Classification within The Common European Framework of Reference for Languages: Learning, Teaching, Assessment: Reading and Listening: Final Report of The Dutch CEF Construct Project. Available at: http://eprints.lancs.ac.uk/44/1/final_report.pdf (Accessed: 10 May 2014).

Alderson, J. C. and Kremmel, B. (2013) Re-examining the content validation of a grammar test: The (im)possibility of distinguishing vocabulary and structural knowledge. *Language Testing*, 30(4), pp.535-556.

Al-Hinai, N. S. (2011) *Effective college teaching and students ratings of teachers: what students think, what faculty believe, and what actual ratings show implications for policy and practice in teaching quality assurance and control in higher education in Oman*. PhD Thesis, Durham University. Available at: http://etheses.dur.ac.uk/649/ (Accessed: 20 September 2015).

Al-Husseini, S. (2004) *An analysis of the English needs of Omani students on vocational and technical courses with implications for the design of foundation year English language programmes.* PhD thesis, the University of Leeds.

Al-Husseini, S. (2006) The visible and invisible role of English foundation programme: A search for communication opportunities within EFL contexts. *The Asian EFL Journal Quarterly*, 8(4), pp. 35-51. Available at: http://www.asian-efl-journal.com/December_2006_EBook.pdf (Accessed: September 2015).

Al-Issa, A. S.M. (2002) *An ideological and discursive analysis of English language teaching in the Sultanate of Oman*. PhD thesis, University of Queensland, Australia.

Al-Issa, A. S.M. (2005) An ideological discussion of the impact of the NNESTs' English language knowledge on ESL policy implementation: A special reference to the Omani context. *The Asian EFL Journal Quarterly*,

7(3), pp.98-112. Available at: http://www.asian-efl-journal.com/September_2005_EBook_editions.pdf (Accessed: 16 September 2015).

Al-Issa, A. S.M. (2006) *The cultural and economic politics of English language teaching in Sultanate of Oman*. *Asian EFL Journal Quarterly, 8(1)*. Available *at:* http://asian-efl-journal.com/1144/quarterly-journal/2006/03/the-cultural-and-economic-politics-of-english-language-teaching-in-sultanate-of-oman/ *(*Accessed: 20 September 2015).

American Psychological Association (2010) *Publication manual of the American psychological association*. Washington, DC: American Psychological Association.

Anderson, R. C. and Pearson, P. D. (1988) A schema-theoretic view of basic processes in reading comprehension. In: Carrell, P. L., Devine, J. and Eskey, D. E. (eds.) *Interactive approaches to second language reading*. Cambridge: Cambridge University Press. pp.37-55.

Anderson, N. J. (1999a) Improving reading speed: activities for classroom. English Teaching Forum, 37(2). pp.2-5. Available at: http://dosfan.lib.uic.edu/usia/E-USIA/forum/vols/vol37/no2/p2.htm (Accessed: 21 July 2013).

Anderson, N. J. (1999b) *Exploring second language reading: issues and strategies*. London: Heinle & Heinle Publishers.

Antonius, R. (2003) *Interpreting quantitative data with SPSS*. London, Thousand Oaks, New Delhi: Sage Publications.

Ardasheva, Y., Tretter, T. R. and Kinny, M. (2012) English Language Learners and academic achievement: revisiting the threshold hypothesis.

Language Learning: A Journal of Research in Language Studies, 62(3), pp.769–812.

Athey, I. J. (1971) Language models and reading. *Reading Research Quarterly*, 7(1), pp.16-110.

Bachman, L. F. (1990) *Fundamental considerations in language testing*. Oxford: Oxford University Press.

Bachman, L. F. (2000) Modern testing at the turn of the country: assuring that what we count counts. *Language Testing*, 17(1), pp.1-42.

Bachman, L. F. (2004) *Statistical analysis for language assessment*. Cambridge: Cambridge University Press.

Bachman, L., F. and Palmer, A. S. (1996) *Language Testing in Practice*. Oxford: Oxford University Press.

Bachman, L. F., Davidson, F. and Milanovic, M. (1996) The use of test method characteristics in the content analysis and design of EFL proficiency tests. Language Testing, 13(2), pp.125-150.

Bachman, L. F. and Palmer, A. S. (2010) Language assessment in practice. Oxford: Oxford University Press.

Bax, S. (2013) The cognitive processing of candidates during reading tests: evidence from eye-tracking. *Language Testing*, 30(4), pp.441-465.

Bell, J. (2010) *Doing your research project: a guide for first-time researchers in education, health and social science*. Maidenhead: McGraw-Hill Open University Press.

Benati, A. G. (2009) *Issues in second language proficiency*. London; New York: Continuum.

Bernhardt, E. B. (1991) *Reading development in a second language: theoretical, empirical and classroom perspectives*. Norwood, NJ: Ablex.

Brindley, G. (2001) Assessment. In: Carter, R. and Nunan, D. (eds.) *The Cambridge Guide to teaching English to speakers of other languages*. Cambridge: Cambridge University Press. pp.137-143.

Brown, J. D. (2001) Using surveys in language programs. Cambridge: Cambridge University Press.

Brown, J. D. (2012) Choosing the right type of assessment. In: Coombe, C., Davidson, P., O'Sullivan, B. and Stoynoff, S. (eds.) *The Cambridge guide to second language assessment*. Cambridge: Cambridge University Press. pp.133-139.

Brown, A. and McNamara, T. (1992) The role of test-taker feedback in the test development process: test taker's reactions to a tape-mediated test of proficiency in spoken Japanese. *Language Testing,* 10 (3), pp. 277-301.

Brown, J. D. and Rodgers, T. S. (2002) *Doing second language research*. Oxford: Oxford University Press.

Brown, J. D., Janssen, G., Trace, J. and Kozhevnikova, L. (2012) A preliminary study of cloze procedure as a tool of eliminating English readability for Russian students. *Second Languages Studies*, 31(1), pp.1-22.

Bruder, M. N. and Henderson, R. T. (1985) *Beginning reading in English as a second language*. Washington: CAL center for Applied Linguistics. pp. Available at*:* http://files.eric.ed.gov/fulltext/ED271940.pdf (Accessed: 17 June 2013).

Bryman, A. (2001) *Social research methods*. Oxford: Oxford University Press.

CaMLA (2015) *MELAB*. Available at: http://www.cambridgemichigan.org/resources/melab/reports (Accessed: 19 September 2015).

Canale, M. (1983) From communicative competence to communicative language pedagogy. In: Richards, J. C. and Schmidt, R. (eds.) *Language and communication*. London and New York: Routledge. pp.2-28.

Canale, M. and Swaine, M. (1980) Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics*, 1(1), pp.1-47.

Cappelli, G. (2010) Lexical complexity: theoretical and empirical aspects. In: Pinnavaia L. and Brownlees, N. (eds.) *Insights into English and Germanic lexicology and lexicography: past and present perspectives*. Monza: Polimetrica International Publishers. pp.115-127.

Carrell, P. L. (1984) The effects of rhetorical organization on ESL readers. *TESOL Quarterly*, 18(3), pp.441-469.

Carrell, P. L. (1988) Introduction: Interactive approaches to second language reading. In: Carrell, P. L., Devine, J. and Eskey, D. E. (eds.) *Interactive approaches to second language reading*. Cambridge: Cambridge University Press. pp. 1-7.

Carrell, P. L., Devine, J. and Eskey, D. E. (eds.) (1988) *Interactive approaches to second language reading*. Cambridge: Cambridge University Press.

Carter, R. and McCarthy, M. (2006) *Cambridge grammar of English: a comprehensive guide to spoken and written English usage*. Cambridge: Cambridge University Press.

Carter, R. and Nunan, D. (eds.) (2001) *The Cambridge Guide to teaching English to speakers of other languages*. Cambridge: Cambridge University Press.

Chang, A. C-S. and College, H. (2010) The effects of a timed reading activity on EFL learners: speed, comprehension, and perceptions. *Reading in a Foreign Language*, 22(2), pp.284-303.

Chihara, T., Sakurai, T. and Oller, Jr., J. W. (1989) Background and culture as factors in EFL reading comprehension. *Language Testing*, 6(2), pp.143-151.

CIA Office of Public Affairs (2015) *The world fact book*. Washington D.C. Available at: https://www.cia.gov/library/publications/the-world-factbook/geos/mu.html (Accessed: 18 September 2015).

Clapham, C. (1996) *The developments of IELTS: a study in the effect of background knowledge on reading comprehension, studies in language testing 4*. Cambridge: UCLES/Cambridge University Press.

Cobb, T. (2003) *VocabProfile, the complete lexical tutor*. Available at: http://www.lextutor.ca. (Accessed: 10 March 2014).

Cobb, T. (2013) Range for texts v.3 [computer program]. Available at: http://www.lextutor.ca/cgi-bin/range/texts/index.pl (Accessed: 15 Sept 2013).

Cohen, A. D. (1994) *Assessing language ability in the classroom*. 2nd edn. Boston, Massachusetts: Heinle & Heinle Publishers.

Cohen, A. D. (1998) Strategies and process in test taking and SLA. In: Bachman, Lyle F. and Cohen, Andrew, D. (eds.) *Interfaces between second language acquisition and language testing research*. Cambridge: Cambridge University Press. pp.90-111.

Cohen, A. D. (2000) Exploring strategies in test-taking: fine-tuning verbal reports from respondents. In: Ekbatani, G. and Pierson, H. (eds.) *Learner-directed assessment in ESL*. New Jersey; London: Lawrence Erlbaum Associates Publishers, pp.127-150.

Cohen, A. D. (2012) Test-taking strategies. In: Coombe, C., Davidson, P., O'Sullivan, B. and Stoynoff, S. (eds.) *The Cambridge guide to second language assessment*. Cambridge: Cambridge University Press. pp.96-104.

Cohen, A. D. and Olshtain, E. (1993) The production of speech acts by EFL learners. *TESOL QUARTERLY,* 27(1), pp.33-56.

Cohen, L., Manion, L. and Morrison K. (2011) *Research methods in education*. 7th edn. London: Routledge.

Cohen, A. D. and Upton, T. A. (2014) I want to go back to the text: response strategies on the reading subtest of the new TOEFL. *Language Testing,* 24(2), pp.209-250.

Coombe, C., Davidson, P., O'Sullivan, B. and Stoynoff, S. (eds.) (2012) *The Cambridge guide to second language assessment*. Cambridge: Cambridge University Press.

Cooper, M. (1984) Linguistic competence of practised and unpractised non-native readers of English. In: Alderson, J. Charles and Urquhart, A. H. (eds.) *Reading in a foreign language*. London; New York: Longman. pp.122-135.

Courtis, S. A. (1914) Standard tests in English. *The Elementary school Teacher*, 14(8), pp.374-392.

Coxhead, A. (2000) A new academic word list. *Tesol Quarterly*, 34 (2), pp. 213-238.

Creswell, J. W. (2003) *Research design: qualitative, quantitative, and mixed methods approaches*. 2nd ed. London: Sage.

Creswell, J. W. (2009) *Research design: qualitative, quantitative, and mixed methods approaches.* 3rd edn. London: Sage.

Crystal, D. (1996) *Discover grammar*. Harlow: Longman.

Csapó, B. and Nikolov, M. (2009) The cognitive contribution to the development of proficiency in a foreign language. *Learning and Individual Differences*, 19(2), pp.209–218.

Cumming, A. (2008) Assessing oral literacy and literate ability. In: Shohamy, E., and Hornberger, Nancy H., (eds.) *Encyclopaedia of language and education: Language Testing and Assessment*, *Volume 7*. New York: Springer. pp.3-18.

Cushing-Weigle, S. (2000) The Michigan English language assessment battery (MELAB): test review. *Language Testing*, 17(4), pp.449-455.

Cushing-Weigle (2002) *Assessing writing*. Cambridge: Cambridge University Press.

Cushing-Weigle, S. and Jensen, L. (1996) Reading rate improvement in university ESL classes. *CATESOL Journal*, 9(2), pp.55-71.

Davies, A. (ed.) (1968) *Language testing symposium: a psycholinguistic approach*. London: Oxford University Press.

Davies, A. (1981) Communicative Syllabus Design by John Munby Review. *TESOL Quarterly*, 15(3), pp.332-336.

Davies, A. (1990) *Principles of language testing*. Oxford: Basil Blackwell.

Davies, A. (2011) Kane, validity and soundness. *Language Testing*, 29(1), pp.37-42.

Day, R. and Bamford, J. (1998) *Extensive reading in the second language classroom*. Cambridge: Cambridge University Press.

Degand, L. and Sanders, T. (2002) The impact of relational markers on expository text comprehension in L1 and L2. *Reading and Writing: An Interdisciplinary Journal* 15(7), pp.739–757.

Derrida. J. (2004) *Positions*. London: Continuum.

Douglas, D. (2009) *Understanding language testing*. London: Hodder Education.

Ekbatani, G. and Pierson, H. (eds.) (2000) *Learner-directed assessment in ESL*. New Jersey; London: Lawrence Erlbaum Associates Publishers.

Embretson, S. (1983) Construct validity: construct representation versus nomothetic span. *Psychological Bulletin*, 93(1), pp.179-197.

Enright, M., Grabe, W., Koda, K., Mosenthal, P., & Mulcahy-Ernt, P. (2000) *Toefl 2000 reading framework: a working paper*. Princeton, NJ: Educational Tesing Service. [TOEFL Monograph Series MS – 17.]

Ericsson, K. A. (2006) Protocol analysis and expert thought: Concurrent verbalizations of thinking during experts' performance on representative tasks. In: Ericsson, K. A., Charness, N., Feltovich, P. J., and Hoffman, R. R. (eds.) *The Cambridge handbook of expertise and expert performance*, pp.223-242.

Ericsson, K. A., Charness, N., Feltovich, P. J. and Hoffman, R. R. (eds.) (2006) *The Cambridge handbook of expertise and expert performance*. Cambridge: Cambridge University Press.

ETS (2009) *Guidelines for the assessment of English language learners*. Available at: https://www.ets.org/s/about/pdf/ell_guidelines.pdf (Accessed: 20 July 2013).

ETS (2014) *Standards for Quality and fairness.* Available from: https://www.ets.org/s/about/pdf/standards.pdf (Accessed: 13 February 2015).

ETS (2015) *TOEFL research program*. Available at: https://www.ets.org/toefl/research/ (Accessed: 19 September 2015).

Farhady, H. (2012) Principles of language assessment. In: Coombe, C., Davidson, P., O'Sullivan, B. and Stoynoff, S. (eds.) *The Cambridge guide to second language assessment*. Cambridge: Cambridge University Press. pp.37-46.

Farrall, M. L. (2012) *Reading assessment: linking language, literacy, and cognition*. Somerset, NJ: Wiley.

Favreau, M. and Segalowitz, N. S. (1982) Second language reading in fluent bilinguals. *Applied Psycholinguistics*, 3(4), pp.329-341.

Field, A. (2001) *Discovering statistics using SPSS*. 2nd edn. London: Sage.

Field, A. (2012) *Discovering statistics using SPSS*.4th Ed. London; Thousand Oaks; New Delhi: Sage Publications.

Flowerdew, J. and Miller, L. (2012) Assessing listening. In: Coombe, C., Davidson, P., O'Sullivan, B. and Stoynoff, S. (eds.) *The Cambridge guide to second language assessment*. Cambridge: Cambridge University Press. pp.225-233.

Foster, J. J. (2001) *Data analysis using SPSS for Windows*. London: Sage.

Frank, H. and Althoen, Steven C. (1994) *Statistics: concepts and applications*. Cambridge: Cambridge University Press.

Freedle, R. (1997) The relevance of multiple-choice reading test data in studying expository passage comprehension: the saga of a 15 year effort towards an experimental/correlational merger. *Discourse Processes*, 23(3), pp.399-440.

Freedle, R. and Kostin, I. (1994) Published can multiple-choice reading tests be construct-valid? A reply to Katz, Lautenschlager, Blackburn, and Harris. *Psychological Science*, 5(2), pp.107-110.

Fulcher, G. (1998) Widdowson's model of communicative competence and the testing of reading: an exploratory study. *System*, 26(3). pp.281-302.

Fulcher, G. (1999) Assessment in English for academic purposes: putting content validity in its place. *Applied linguistics*, 20(2), pp.221-231.

Fulcher, G. (2010) *The reification of the Common European Framework of Reference (CEFR) and effect-driven testing*. In: Psaltou-Jocey, A. (ed.) *Advances in research on language acquisition and teaching: selected papers*. pp.15-26. Available at: http://www.enl.auth.gr/gala/14th/Papers/Invited%20Speakers/Fulcher.pdf (Accessed: 20 July 2013).

Fulcher, G. and Davidson, F. (2007) *Language testing and assessment: an advanced resource book.* Abingdon: Routledge.

Gilbert, N. (2008) *Researching social life.* 3rd edn. London: Sage Publications.

Gillham, B. (2000) *Developing a questionnaire*. London; New York: Continuum.

Goh, S. T. (1990) The effects of rhetorical organization in expository prose on ESL readers in Singapore. *RELC Journal*, 21(2), pp.1-13.

Goldman, S. R. and Rakestraw, J. A. (2000) Structural aspects of constructing meaning from text. In: Kamil, M. L., Mosenthal, P. B., Pearson, P. D., & Barr, R. (eds.) *Handbook of reading research*, Vol. 3. Mahwah, NJ: Lawrence Erlbaum. pp.311-336.

Goodman, K. (1967) Reading: A psycholinguistic game. *Journal of the reading specialist*, 6(4), pp.126-135.

Goodman, K. (1988) The reading process. In: Carrell, P. L., Devine, J. and Eskey, David, E. (eds.) *Interactive approaches to second language reading*. Cambridge: Cambridge University Press. pp.11-21.

Goulden, R., Nation, P. and Read, J. (1990) How large can a receptive vocabulary be? *Applied linguistics*, 11(4), pp.341-363.

Grabe, W. (1988) Reassessing the term “interactive”. In: Carrell, P. L., Devine, J. and Eskey, David, E. (eds.) *Interactive approaches to second language reading*. Cambridge: Cambridge University Press. pp.56-72.

Grabe, W. (1991) Current development in second language research. *Tesol Quarterly*, 25(3), pp.378-388.

Grabe, W. (2009) *Reading in a second language: moving from theory to practice*. New York, Cambridge University Press.

Grabe, W. and Stoller, F. L. (2002) *Teaching and researching reading*. Longman: London.

Graesser, A., McNamara, D. S., Louwerse, M. and Cai, Z. (2004) Coh-Metrix: Analysis of text on cohesion and language. *Behavioral Research Methods, Instruments, and Computers*, 36(2), pp.193-202.

Greasley, P. (2008) *Quantitative data analysis using SPSS: an introduction for health & social science*. Maidenhead: Open University Press.

Green, A. (1998) *Verbal protocol analysis in language testing research: A handbook* (Vol. 5). Cambridge: Cambridge University Press.

Green, A. (2011) A case of testing L2 English reading for class level placement. In: O'Sullivan, B. (ed.) *Language testing: theories and practices*. Basingstoke; New York: Palgrave Macmillan. pp.186-207.

Green, A. (2014) *Exploring language assessment and testing: language in action*. London and New York: Routledge.

Green, A., Unaldi, A., & Weir, C., (2010) Empiricism versus connoisseurship: Establishing the appropriacy of texts in tests of academic reading, Language Testing, 27 (2), pp191-211**.**

Green, A. and Hawkey, R. (2011) Re-fitting for a different purpose: a case study of item writer practices in adapting source texts for a test of academic reading. *Language Testing*, 29(1), pp.109–129.

Green, A., Khalifa, H. and Weir, C. J. (2013) *Examining textual features of reading texts – a practical approach. Cambridge Research Notes, Issue 52*. Cambridge: Cambridge English Language Assessment. Available at: http://www.cambridgeenglish.org/images/139525-research-notes-52-document.pdf (Accessed: 11 March 2014).

Green, N. (2008) Formulating and refining a research question. In: Gilbert, N. (ed.) *Researching social life*. 3rd edn. London: Sage Publications. pp.43-62.

Green, T. and Jay, D. (2005) *Quality assurance and quality control: reviewing and pretesting examination material at Cambridge ESOL*. Available at: http://www.cambridgeenglish.org/images/23140-research-notes-21.pdf (Accessed: 5 March 2014).

Groot, P. J. M. (1975) Validation of language tests. In: Palmer, L. and Spolsky, B. (eds.) *Papers on language testing 1967-1974*. Washington: Teachers of English To Speakers of Other Languages. pp.128-137.

Gudykunst, W. B. (2004) *Bridging differences: effective intergroup communication*. 4th edn. Thousand Oaks, Calif.: Sage Publications.

Hall, C. J. and Durán, A. R. (2009) Cross-Linguistic influences in L2 verb frames: the effects of word familiarity and language proficiency. In: Benati, Alessandro G. (ed.) *Issues in second language proficiency*. London; New York: Continuum. pp.24-44.

Harris, D. P. (1968) The linguistics of language testing: In Davies, A. (ed.) *Language testing symposium: a psycholinguistic approach*. London: Oxford University Press. pp. 36-45.

Hayes-Harb, R. (2006) Native speakers of Arabic and ESL texts: evidence for the transfer of written word identification processes. *TESOL Quarterly*, 40(2), pp.321-339.

Henriksen, B. (1999) Three dimensions of vocabulary development. *Studies in Second Language Acquisition*, 21(2), pp.303–317.

Higher Education Admissions Centre (2015) *Summary of admission statistics*. Higher Muscat: Ministry of Higher Education. Available at:

http://www.heac.gov.om/index.php?option=com_content&view=article&id=56&Itemid=67&lang=en (Accessed: 21 September 2015).

Hill, C. and Parry, K. (1992) The test at the gate: models of literacy in reading assessment. *TESOL Quarterly*, 26(3) pp.433-461.

Hubley, N. J. (2012) Assessing reading. In: Coombe, C., Davidson, P., O'Sullivan, B. and Stoynoff, S. (eds.) *The Cambridge guide to second language assessment*. Cambridge: Cambridge University Press. pp.211-217.

Huckin, T., Haynes, M. and Coady, J. (1993) *Second language reading and vocabulary learning*. New Jersey: Ablex Publishing Corporation.

Huddleston, R. and Pullum, G. K. (2005) *A student's introduction to English grammar.* Cambridge: Cambridge University Press.

Hudson, T. (2005) Trends in assessment scales and criterion-referenced language assessment. *Annual Review of Applied Linguistics*, 25, pp.205-227.

Hudson, T. (2007) *Teaching second language reading*. Oxford: Oxford University Press.

Hughes, A. (2003) *Testing for language teachers*. Cambridge: Cambridge University Press.

Hymes, D. (1972) On communicative competence. In: Pride, J. B. and Holmes, J. (eds.) *Sociolinguistics*. Harmondsworth and New York: Penguin, pp. 269-293.

Ibra College of Technology (2014a) *Engineering department – programs offered*. Available at: http://www.ict.edu.om/ (Accessed: 23 March 2014).

Ibra College of Technology (2014b) *Business department – programs offered*. Available at: http://www.ict.edu.om/ (Accessed: 23 March 2014).

Ibra College of Technology (2014c) *Information technology department – programs offered*. Available at: http://www.ict.edu.om/ (Accessed: 23 March 2014).

IELTS (2015) *Research: IELTS jointly funded and published research*. Available at: http://www.ielts.org/researchers/research.aspx (Accessed: 19 September 2015).

IHS Global Insight Inc. (2012) *Country intelligence: reports – Oman*. Available at: www.brad.ac.uk (Accessed: 19 September 2015).

IHS Global Insight Inc. (2015) *Country reports – Oman: 31 Aug 2015 IHS Economics and country risk*. Available at: http://web.b.ebscohost.com/ehost/pdfviewer/pdfviewer?sid=4e3f7738-06e3-4223-9045-b524cfec7828%40sessionmgr111&vid=15&hid=123 (Accessed: 21 September 2015).

Kamil, M. L., Mosenthal, P. B., Pearson, P. D., & Barr, R. (eds.) (2000) *Handbook of reading research*, Vol. 3. Mahwah, NJ: Lawrence Erlbaum

Kane, M. (2002) Validating high-stakes testing programs. *Educational Measurement: Issues and Practice*, 21(1), pp.31-41.

Kantarcioğlu, E. (2012) *Relating an institutional proficiency examination to the CEFR: a case study*. PhD Thesis, University of Roehampton. Available at: http://roehampton.openrepository.com/roehampton/bitstream/10142/278935/1/Elif%20Kantarcioglu%20PhD%202012.pdf (Accessed: 20 September 2015).

Kaplan, R. B. and Grabe, W. (2002) A modern history of discourse analysis. *Journal of Second Language Writing*, 11(3), pp.191-223.

Kasper, G. and Kellerman, E. (eds.) (1997) *Communication strategies: Psycholinguistic and sociolinguistic perspectives*. London and New York: Routledge.

Kauffman, D. F. and Kiewra, K. A. (2009) What makes a matrix so effective? An empirical test of the relative benefits of signaling, extraction, and localization. *Instructional Science*, 38(6). pp.679–705.

Khalifa, H. (2005) *Are test taker characteristics accounted for in Main Suite Reading papers? Research notes 21*. Available at: http://www.cambridgeesol.org/rs_notes/rs_nts21.pdf (Accessed: 7 April 2014).

Khalifa, H. and Weir, C. (2009) *Examining reading: research and practice in assessing second language reading*. Cambridge: Cambridge University Press.

Kintsch, W. and Yarbrough, J. C. (1982) Role of rhetorical structure in text comprehension. *Journal of Educational Psychology*, 74(6), pp.828-834.

Kirkland, M. R. and Saunders, M. A. P. (1991) Maximizing student performance in summary writing: managing cognitive load. *TESOL Quarterly*, 25(1), pp.105-121.

Klusewitz, M. A. and Lorch, JR. R. F. (2000) Effects of headings and familiarity with a text on strategies for searching a text. *Memory and Cognition*, 28 (4), pp.667-676.

Kobayashi, M. (1993) Method effects on reading comprehension test performance: text organization and response format. *Language Testing*, 19(2), pp.193-220.

Koda, K. (2004) *Insights into second language reading: a cross-linguistic approach*. Cambridge: Cambridge University Press.

Kunnan, A. (1995) *Test taker characteristics and test performance: A structural modelling approach*. Cambridge: Cambridge University Press.

Kunnan, A. (2004) Test fairness. In: Milanovic, M. and Weir, Cyril W. (eds.) *European language testing in a global context: selected papers from the ALTE conference, Barcelona*. Cambridge: Cambridge University Press. pp.27–48.

Laberge, D. and Samuels, S. J. (1974) Toward a theory of automatic information processing in reading. *Cognitive Psychology*, 6(2), pp.293-323.

Lacroix, N. (1999) Macrostructure construction and organization in the processing of multiple text passages. *Instructional Science*, 27(3), pp.221-233.

Lado, R. (1961) *Language testing: the construction and use of foreign language tests.* London: Longman.

Larkin, J. H. and Simon, H. A. (1987) Why a diagram is (sometimes) worth ten thousand words. *Cognitive Science*, 11(1). pp.65–99.

Laufer, B. and Nation, P. (1995) Vocabulary size and use: Lexical richness in L2 written productions. *Applied Linguistics*, 16 (3), pp.307-322.

Leech, G. and Svartvik, J. (2002*) A communicative grammar of English*. 3rd edn. London: Longman.

Lems, K., Miller, L. D. and Soro, T. M. (2010) *Teaching reading to English language learners*. New York; London: The Guildford Press.

Leow, R.P. and Morgan-Short, K. (2004) To think aloud or not to think aloud: the issue of reactivity in SLA research methodology. *Studies in Second Language Acquisition*, 26(1), pp.35-37.

Malone, M. E. (2010) Canadian academic English language (CAEL) assessment. *Language Testing*, 27(4), pp.631–636.

McKay, S. (1980) Communicative Syllabus Design. *The Modern Language Journal*, 64(1), pp.151-152.

McNamara, D. S., Louwerse, M. M., Cai, Z. and Graesser, A. (2005) *Coh-Metrix version 1.4*. Available at: http://141.225.42.101/cohmetrix3/signinup.aspx (Accessed: 10 March 2014).

McNamara, D. S. and Graesser, A. C. (2012) Coh-Metrix: an automated tool for theoretical and applied natural language processing. In: McCarthy, Philip M. and Roonthum-Denecke, C. (eds.) *Applied natural language processing: identification, investigation and resolution*. Hershey PA: Information Science Reference. pp.188-205.

McNamara, T. (2000) *Language testing*. Oxford: Oxford University Press.

Mead, R. (1982) Review of Munby: communicative syllabus design. *Applied Linguistics*, 3(1), pp.70-78.

Messick, S. (1989) Meaning and values in test validation: the science and ethics of assessment. *Educational Researcher*, 18(2), pp. 5-11.

Messick, S. (1995) Standards of validity and the validity of standards in performance assessment. *Educational Measurement: Issues and Practice*, 14(4), 5-8.

Milanovic, M. and Weir, Cyril W. (eds.) (2004) *European language testing in a global context: selected papers from the ALTE conference, Barcelona*. Cambridge: Cambridge University Press. pp.27–48.

Miller, D. C. and Salkind, N. J. (2002) *Handbook of research design and social measurement*. 6th edn. Thousand Oaks, Calif.; London: Sage Publications.

Ministry of Manpower (2004) *Ministerial decision no.72/2004: Colleges of Technology bylaws*. Muscat: Ministry of Manpower.

Moore, T. (2002) Knowledge and agency: a study of 'metaphenomenal discourse' in textbooks from three disciplines. *English for Specific Purposes*, 21(4), pp.347–366.

Moore, T. and Morton, J. (1999) *Authenticity in the IELTS academic module writing test: a comparative study of Task 2 items and university assignments in IELTS research reports*. Volume 2, IELTS Australia: Canberra, pp 64-106.

Moore, T. and Morton, J. (2005) Dimensions of difference: a comparison of university writing and IELTS writing. *Journal of English for Academic Purposes*, 4(1), pp.43–66.

Moore, T., Morton, J. and Price, S. (2011) *Construct validity in the IELTS academic reading test: a comparison of reading requirements in IELTS test items and in university study.* Available at: http://www.ielts.org/pdf/Vol11_Report_4_Contruct%20validity%20in%20the%20IELTS%20Academic%20Test.pdf (Accessed: 10 April 2014).

Mosback, G. and Mosback, V. (1976) *Practical faster reading*. Cambridge: Cambridge University Press.

Moustakas, C. (1994) *Phenomenological research methods*. California: Sage Publications, Thousand Oaks.

Munby, J. (1978) *Communicative syllabus design: a sociolinguistic model for defining the content of purpose-specific language programmes*. Cambridge: Cambridge University Press.

Murray, J. A. H. and others (eds.) (1933) *The Oxford English dictionary: being a corrected re-issue with an introduction, supplement, and bibliography of a new English dictionary on historical principles, founded mainly on the materials collected by the Philological Society*. VOLUME III D-E. Oxford: The Clarendon Press.

Muscat Media Group (2014) *Business: Oman's diversification strategy bearing fruit. Times of Oman* [Online]. Available at: http://timesofoman.com/article/39808/Business/Oman's-diversification-strategy-bearing-fruit (Accessed: 19 September 2015).

Mustafa, Z. (1995) The effect of genre awareness on linguistic transfer. *English for Specific Purposes*, 14(3), pp.247- 256.

Nardi, Peter M. (2003) *Doing survey research: a guide to quantitative methods.* Boston, Mass; London: Allyn and Bacon.

Nation, P. (2005) Reading faster. *PASAA*, 36, pp.21-37.

Norris, J. M., Brown, J. D., Hudson, T. and Yoshioka, J. (1998) *Designing second language performance assessments*. Hawai'i: University of Hawai'i.

Norris, S. (1990) Effect of eliciting verbal reports of thinking on critical thinking performance. *Journal of Educational Measurement*, 27(1), pp.41-58.

North, B., (2011) Describing language levels. In O'Sullivan, B.,(ed.) *Language testing: theories and practices*. Hampshire: Palgrave Macmillan. pp.33-59.

Norton, B. and Stein, P. (1998) On why the “Monkeys Passage” bombed: tests, genres and teaching. In: Kunnan, A. J. (ed.) *Validation in language assessment*. Mahwah, New Jersey: Lawrence Erlbaum Associates, Inc, Publishers. pp.231-249.

Nuttall, C. (1996) *Teaching reading skills in a foreign language.* London: Heinemann.

O’Sullivan, B. (2011) (ed.) *Language testing: theories and practices*. Basingstoke; New York: Palgrave Macmillan.

O’Sullivan, B. (2012) The assessment development process. In: Coombe, C., Davidson, P., O’Sullivan, B. and Stoynoff, S. (eds.) *The Cambridge guide to second language assessment*. Cambridge: Cambridge University Press. pp.47-58.

O’Sullivan, B. and Weir, C. J. (2011) Test development and validation. In: O’Sullivan, B. (ed.) *Language testing: theories and practices*. Basingstoke; New York: Palgrave Macmillan. pp.13-32.

OAAA (2008) *Oman academic standards for general foundation programs*. Muscat: Oman Accreditation Council; Ministry of Higher Education. Available at: http://www.oaaa.gov.om/Docs/GFP%20Standards%20FINAL.pdf (Accessed: 21 September 2015).

Office for National Statistics (2015) *Ageing of the UK population*. Available at: http://www.ons.gov.uk/ons/rel/pop-estimate/population-estimates-for-uk--england-and-wales--scotland-and-northern-ireland/mid-2014/sty-ageing-of-the-uk-population.html (Accessed: 18 September 2015).

Oller, J. W. and Tullius, J. R. (1972) Reading skills of non-native speakers of English. *IRAL*, 12(11), pp.69-79.

Olson, D. R. and Torrance, N. (eds.) (2009) *The Cambridge handbook of literacy*. Cambridge: Cambridge University Press.

Oppenheim, A. N. (1992) *Questionnaire design, interviewing and attitude measurement*. London; New York: Pinter.

Pallant, J. (2010) *SPSS survival manual: a step by step guide data analysis using SPSS*. 4th edn. Berkshire: McGraw-Hill.

Palmer, B. C., El-Ashry, F., Leclere, J. T., & Chang, S. (2007) Learning from Abdallah: a case study of an Arabic-speaking child in a U.S. school. *The Reading Teacher*, 61(1), pp.8-17.

Palmer, L. and Spolsky, B. (eds.) (1975) *Papers on language testing 1967-1974*. Washington: Teachers of English to Speakers of Other Languages.

Perfetti, C. A. (1985) *Reading ability*. New York: Oxford University Press.

Pilliner, A. E. C. (1968) Subjective and objective testing. In: Davies, A. *Language testing symposium: a psycholinguistic approach*. London: Oxford University Press. pp.19-35.

Pinnavaia L. and Brownlees, N. (eds.) (2010) *Insights into English and Germanic lexicology and lexicography: past and present perspectives*. Monza: Polimetrica International Publishers.

Pressley, M. and Afflerbach, P. (1995) *Verbal protocols of reading: the nature of constructively responsive reading*. New Jersey; Hove: Lawrence Erlbaum Associates Publishers.

Pride, J. B. and Holmes, J. (eds.) (1972) *Sociolinguistics*. Harmondsworth and New York: Penguin, pp. 269-293.

Procter, M. (2008) Measuring attitudes. In: Gilbert, N. (ed.) *Researching social life*. London: Sage Publications Ltd. pp.206-225.

Purpura, J. E. (1999) *Learner strategy use and performance on language tests: a structural equation modeling approach: studies in language testing 8*. Cambridge; University of Cambridge Local Examinations Syndicate: Cambridge University Press.

Purpura, J. E. (2004) *Assessing grammar*. Cambridge: Cambridge University Press.

Quirk, R., Greenbaum, S., Leech, G., and Svartvik, J. (1985) *A comprehensive grammar of the English language*. London: Longman.

Randall, M. (2009) Second language reading proficiency and word recognition: the concept of salience and its application across different scripts. In: Benati, Alessandro G. (ed.) *Issues in second language proficiency*. London; New York: Continuum. pp.116-131.

Raymond, P. and Parks, S. (2002) Transitions: orienting to reading and writing assignments in EAP and MBA contexts. *Canadian Modern Language Review*, 59(1), pp.152-180.

Read, J. and Chapelle, C. A. (2001) A framework for second language vocabulary assessment*. Language Testing*, 18(1), pp.1–32.

Richards, J. C. and Schmidt, R. (eds.) (1983) *Language and communication*. London and New York: Routledge.

Rimmer, W. (2006) Measuring grammatical complexity: the Gordian knot. *Language Testing*, 23(4), pp. 479-519.

Robinson, D., Katayama, A., DuBois, N., and DeVaney, T. (1998) Interactive effects of graphic organizers and delayed review on concept acquisition. *The Journal of Experimental Education*, 67(1), pp.17–31.

Robson, C. (2011*) Real world research*. 2nd edn. Oxford: Blackwell.

Ross, S. (1997) An introspective analysis of listener inferencing on a second language listening test. In: Kasper, G. and Kellerman, E. (eds.) *Communication strategies: Psycholinguistic and sociolinguistic perspectives*. London and New York: Routledge. pp. 216-237.

Ruddell, R. B. and Unrau, N. J. (eds.) (2004) *Theoretical models and processes of reading*. 5th edn. Newark, DE: International Reading Association.

Rumelhart, D. E. (1977) Toward an interactive model of reading. In: Dornic, S. (Ed.) *Attention and performance VI*. Hillsdale, NJ: Erlbaum. pp.573-603.

Rupp, A., A., Ferne, T. and Choi, H. (2006) How assessing reading comprehension with multiple-choice questions shapes the construct: a cognitive processing perspective. *Language Testing*, 23(4), pp.441-74.

Saigh, K. and Schmitt, N. (2012) Difficulties with vocabulary word form: the case of Arabic ESL learners. *System*, 40(1), pp.24-36.

Samuels, S. J. (2004) Toward a theory of automatic information processing in reading, revisited. In: Ruddell, R. B. and Unrau, N. J. (eds.) *Theoretical models and processes of reading*. 5th edn. Newark, DE: International Reading Association. pp.1127-1148.

Samuels, S. J. and Kamil, M. L. (1988) Models of the reading process. In: Carrell, P. L., Devine, J. and Eskey, D. E. (eds.) *Interactive approaches to second language reading*. Cambridge: Cambridge University Press. pp.22-36.

Sarantakos, S. (2007) *A tool kit for quantitative data analysis using SPSS*. Basingstoke: Palgrave Macmillan.

Sasaki, M. (2000) Effects of cultural schemata on students' test-taking processes for cloze tests: a multiple data source approach. *Language Testing*, 17(1), pp.85–114.

Saville, N. (2000) *Developing language learning questionnaires LLQs: Research notes 1.* Available at: http://www.cambridgeenglish.org/images/23112-research-notes-01.pdf (Accessed: 6 May 2014).

Scott, M. (2006) *Oxford WordSmith tools 4.0*. Available at: http://www.lexically.net/downloads/version4/html/index.html (Accessed: 10 March 2014).

Scott, M. (2014) *Introduction to WordSmith Tools*. Available at: http://www.lexically.net/downloads/version6/HTML/index.html?getting_started.htm (Accessed: 18 March 2014).

Seliger, H. W. and Shohamy, E. G. (1989) *Second language research methods*. Oxford: Oxford University Press.

Shiotsu, T. (2003) *Linguistic knowledge and processing efficiency as predictors of L2 reading ability: a component skills analysis.* PhD Thesis. University of Reading.

Shiotsu, T. and Weir, C. J. (2007) The relative significance of syntactic knowledge and vocabulary breadth in the prediction of second language reading comprehension test performance. *Language Testing*, 24(1), pp.99-128.

Shohamy, E., and Hornberger, N. H., (eds.) (2008) *Encyclopedia of language and education: Language Testing and Assessment*, *Volume 7*. New York: Springer.

Simmons, R. (2008) Questionnaires. In: Gilbert, N. (ed.) *Researching social life*. London: Sage Publications Ltd, pp.182-205.

Smith, F. (2004) *Understanding reading: a psycholinguistic analysis of reading and learning to read*. 4th edn. Mahwah; New Jersey; London: Lawrence Erlbaum Associate Publishers.

Snow, C. (2002) *Reading for understanding: toward a research and development program in reading comprehension*. Pittsburgh: RAND.

Spack, R. (1997) The acquisition of academic literacy in a second language: A longitudinal case study. *Written Communication,* 14(1), pp.3–62.

Special Economic Zone Authority Duqm (2013) *Special Economic Zone Authority*. Special Economic Zone Authority [Online]. Available at: http://www.duqm.gov.om/contact-us/contact-details (Accessed: 22 September 2015).

Spiegel, D. L. and Fitzgerald, J. (1990) Textual cohesion and coherence in children's writing revisited. *Research in the Teaching of English*, 24(1), pp.48-66.

Spyridakis, J. H. and Standal, T. (1987) Signals in expository prose: effects on reading comprehension. *Reading Research Quarterly*, 22(3), pp.285-298.

Sture, J. (2010) *Ethics in research projects: some guidance on recognising and addressing ethical issues*. Available at:

http://www.brad.ac.uk/gateway/media/Gateway/Documents/Ethics/GuidanceinResea rchEthics-JS.pdf (Accessed: 20 May 2012).

Swales, J. (1986) A genre-based approach to language across the curriculum. In M. L. Tickoo (ed.) *language across the Curricuhon* (pp.10-22) Singapore: RELC.

Taha, H.Y. (2013) Reading and Spelling in Arabic: Linguistic and Orthographic Complexity. *Theory and Practice in Language Studies*, 3(5), pp.721-727.

Tavakoli, P. (2009) Researching task difficulty: towards understanding L2 proficiency. In: Benati, Alessandro G. (ed.) *Issues in second language proficiency*. London; New York: Continuum. pp.216-232.

Taylor, L. (2004) Testing times: research directions and issues for Cambridge ESOL examinations. *TESOL Quarterly,* 38(1), pp.141-146.

Taylor, L. (2009) Setting language standards for teaching and assessment: a matter of principle, politics, or prejudice? In: Taylor, L. and Weir, Cyril J. (eds.) *Language testing matters: investigating the wider social and educational impact of assessment – Proceedings of the ALTE Cambridge conference, April 2008*. Cambridge: Cambridge University Press. pp.139-157.

Taylor, L. and Weir, Cyril J. (eds.) (2009) *Language testing matters: investigating the wider social and educational impact of assessment – Proceedings of the ALTE Cambridge conference, April 2008*. Cambridge: Cambridge University Press.

Thabane, L., Ma, J., Chu, R., Cheng, J., Ismaila, A., Rios, L.P., Robson, R., Thabane, M., Giangregorio, L., and Goldsmith, C. H. (2010) A tutorial on pilot studies: the what, why and how. *BMC medical research methodology*, *10*(1), pp.1-10.

The CAEL Assessment Office (2015) *Reports and related articles*. Available at: http://www.cael.ca/index.php (Accessed: 19 September 2015).

The LTTC (2015) *GEPT*. Available at: https://www.lttc.ntu.edu.tw/E_LTTC/E_GEPT.htm (Accessed: 19 September 2015).

The Open University (1998) *Linear Statistical Modelling*. Huddersfield: The Charlesworth Group.

The Public Authority for Investment Promotion and Export Development (2013) *Ithraa*. Available at: https://ithraa.om/EN/Pages/Home.aspx (Accessed: 9 February 2013).

Urquhart, A. H. (1984) The effect of rhetorical ordering on readability. In: Alderson, J. Charles and Urquhart, A. H. (eds.) *Reading in a foreign language*. London; New York: Longman. pp.160-175.

Urquhart, S. and Weir, C. (1998) *Reading in a second language: process, product and practice*. New York: Longman.

US State Department Geographer (2012) *Google Earth*. US: US state of department, ORION ME and LeadDog Consulting. Available at: http://www.google.com/earth/index.html (Accessed: 10 May 2012).

Valette, R. M. (1977) *Modern language testing*. New York: Harcourt Brace Jovanovich.

Wagner, W. E. (2010) *Using SPSS for social statistics and research methods*. 2nd edn. Thousand Oaks, Calif.: Pine Forge Press.

Weir, C. J. (1990) *Communicative language testing*. New York; London: Prentice-Hall.

Weir, C. J. (2005) *Language testing and validation: An evidence-based approach*. Basingstoke; New York: Palgrave Macmillan.

Weir, C. J. (2013) The measurement of reading ability 1913-2012. In: Weir, Cyril J., Vidakovic', I. and Galaczi, E. D. (eds.) *Measured constructs: studies in language testing 37*. Cambridge: Cambridge University Press. pp.109-179.

Weir, C., Hawkey, R., Green, A., Unaldi, A., & Devi, S., (2009) *3 The relationship between the academic reading construct as measured by IELTS and the reading experiences of students in their first year of study at a British university*, IELTS RESEARCH REPORTS, VOLUME 9, British Council and IELTS Australia.

Weir, C., Hawkey, R., Green, A., & Devi, S., (2012) *4 The cognitive processes underlying the academic reading construct as measured by IELTS*, IELTS RESEARCH REPORTS, VOLUME 9, British Council and IELTS Australia.

Weir, C. J., Vidakovic', I., and Galaczi, Evelina D. (eds.) (2013) *Measured constructs: studies in language testing 37*. Cambridge: Cambridge University Press.

WIDA Consortium (2015) *ACCESS for ELLs summative assessment: technical reports*. Available at: https://www.wida.us/assessment/access/ (Accessed: 18 September 2015).

Widdowson, H. G. (1983) *Learning purpose and language use*. Oxford: Oxford University Press.

Wilkins, D. A. (1980) Communicative Syllabus Design by J. Munby. *British Journal of Educational Studies*, 28(2), pp.155-157.

Wisker, G. (2008) *The postgraduate research handbook: succeed with your MA, MPhil, EDD and PhD*. Basingstoke: Palgrave Macmillan.

Wu, R. Y. (2011) *Reading component through a critical evaluation on alignment with the common European framework of reference*. PhD Thesis. University of Bedfordshire.

Yamashita, J. (2003) Processes of taking a gap-filling test: comparison of skilled and less skilled EFL readers. *Language Testing*, 20(3), pp.267–293.

Yanagawa, K. (2012) *A partial validation of the contextual validity of the centre listening test in Japan*. PhD Thesis. University of Bedfordshire.

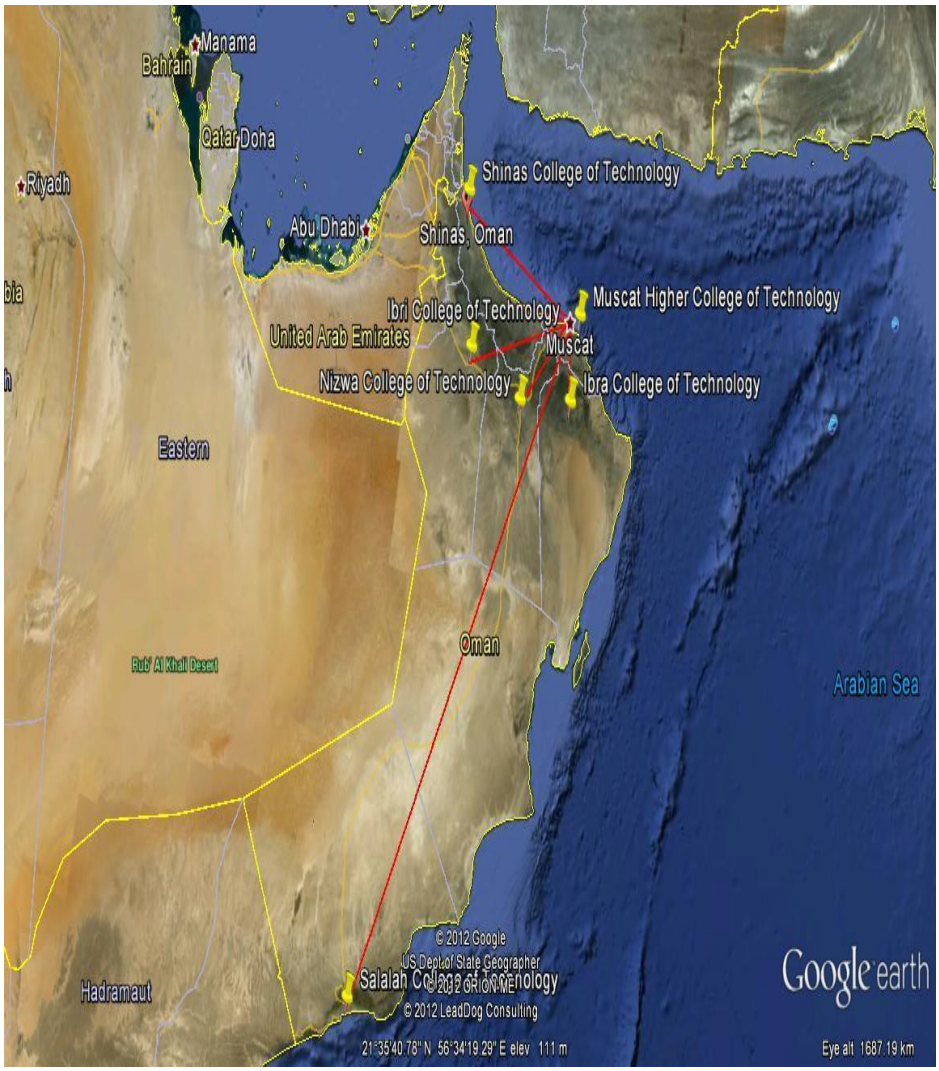
Yoshimura, F. (2000) Automaticity Theory and EFL in Japan: With Some Specific Applications for Reading. *Literacy Across Cultures*, 4(1). Available at: http://www2.aasa.ac.jp/~dcdycus/LAC2000/yoshimura.htm (Accessed: 15 October 2013).

Young, J. W. (2008) *Ensuring valid content tests for English language learners*. Available at: https://www.ets.org/Media/Research/pdf/RD_Connections8.pdf (Accessed: 7 June 2013).

Zareva, A. (2005) Models of lexical knowledge assessment of second language learners of English at higher levels of language proficiency. *System*, 33(4), pp.547–562.

Zheng, Y. and Cheng, L. (2008) College English test (CET) in China: test reviews. *Language Testing*, 25(3), pp.408-417.

Appendix 1: Approximate distance of the seven Colleges of Technology (CTs) from the Capital of Oman, Muscat



Downloaded and reproduced from US State Department Geographer (2012)

College of Technology	Approximate Distance from Capital of Oman, Muscat (km)
Higher College (HCT)	Within Muscat area
Ibra (ICT)	130
Nizwa (NCT)	140
Shinas (SCT)	250
Salalah (SCT)	1,000
Ibri (CT)	300

Appendix 2 Test Task used for the natural experiment via VPA

Read the passage about Zanzibar and answer the questions. (25 marks)

Over the past 1000 years, the group of islands knows as Zanzibar has held an important place in global trade. Today, it remains an attractive destination for tourists and researchers who want to learn more about the unique history and culture of the region.

Zanzibar is located about 30 miles off the coast of Africa. It is actually composed of two main islands as well as numerous small ones. The largest island is called Unguja in Swahili, but commonly known as "Zanzibar Island". The second main island is Pemba, located about 50 kilometers to the north of Unguja. The climate is tropical and humid. It is warm year-round and there are two different rainy seasons: brief showers occur in November and longer rains come in March, April, and May. Though the island was originally thickly covered with forest, now most of its trees are found in Jozani Chawaka Bay National Park. Both the Zanzibar leopard, a kind of large cat, and a red colobus monkey live in this habitat. Around the coast, the sandy beaches are protected by coral reefs, which are full of marine life.

The first permanent settlers in Zanzibar probably arrived in about 1000 AD. Little is known about the early history of the islands, but Zanzibar became an important headquarters for merchants from Persia, Arabia, India, China, and Portugal by the year 1500. Goods such as ivory, spices, glassware, and textiles were traded *extensively*. Because of the prevalence of the spice trade, Zanzibar was known as the Spice Islands. Rich and *prosperous*, Zanzibar caught the attention of other countries. The Portuguese gained control of Zanzibar in 1504. They ruled there until 1698, when the Sultanate of Oman took power over the islands. Under Omani leadership, Zanzibar became the world's largest producer of cloves and the largest slave trading center on the east coast of Africa. In 1832, Said bin Sultan Al-Said, Sultan of Oman, relocated his capital from Muscat to Zanzibar, illustrating the importance of the islands' wealth and resources. Zanzibar became independent of Oman after the Sultan's death in 1856, and soon fell under British control. It wasn't until 1964 that the islands became part of its current country, Tanzania.

It's estimated that 1 million people live in Zanzibar. Most of the population is African in origin, but because of the colorful history of the region, Indians and Arabs live here as well. Life expectancy at birth is 57 years and fish, rice, yams, and tropical fruit make up an important part of the local diet. About a tenth of the population is employed in the fishing industry. Zanzibar also manufactures clove oil and woven goods.

Today, though fishing and farming dominate the economy, the local government wants to promote tourism. Stone Town, the capital of Zanzibar, has recently been declared a UNESCO World Heritage Site. At the moment, only about 100,000 tourists visit Zanzibar every year, but this is expected to grow in the next few years as more and more people learn about this exotic location.

PART A: Circle T if the statement is true, F if it is false. (4 marks)

1) Pemba is an island to the north of Unguja. T F

2) The climate in Zanzibar is mild and dry. T F

3) Zanzibar was called the Spice Islands T F

4) Tourism dominates the Zanzibar economy. T F

PART B: Circle the best answer: (6 marks)

5) What is the main idea of **paragraph 1?**

a) Zanzibar is an island nation off the coast of Africa.

b) Zanzibar is a historically important and interesting place.

c) Zanzibar is full of tourists who want to see the clove trade.

6) What is the main idea of **paragraph 3**?

a) Many things have happened in Zanzibar's long history

b) Little is known about the early history of the islands.

c) The ivory trade was important in the past.

7) What is the main idea of **paragraph 5**?

a) Tourism in Zanzibar is being promoted.

b) Stone Town is very important.

c) UNESCO recently visited Stone Town.

PART C: Circle the best answer: (2 marks)

8) Find the word *extensively* in line 16. Choose the best meaning for this word in the context.

a) not very much b) a lot c) occasionally

9) Find the word *prosperous* in line 17. Choose the best meaning for this word in the context.

a) wealthy b) poor c) dangerous

PART D: Answer the questions. (4 marks)

10) What protects the beaches?

_____

11) What did Zanzibar become the largest producer of?

_____

12) Who moved his capital to Zanzibar?

_____

13) What is the origin of most of the population in Zanzibar?

_____

PART E: Complete the table about the History of Zanzibar. (4 marks)

Year	**Event/Fact**
By 1500	Zanzibar was an important trade headquarters.
1504	**(14)** _____ took control of Zanzibar.
(15)_____	The Sultanate of Oman took power.
1832	The Sultan relocated his capital to Zanzibar
(16)_____	Sultan Said bin Sultan al-Said died.
1964	Zanzibar joined **(17)** _____

PART F: Complete the paragraph with the best words from the box. There are <u>TWO EXTRA</u> words. (5 marks)

destination	global	headquarters	manufactures
population	promoting	unique	

The islands known as Zanzibar have a **(18)** _____and interesting history, culture and environment. Over many centuries Zanzibar, which is also known as the Spice Islands, has been important in **(19)** _____trading. Merchants from many different nations have gone there to trade their goods in exchange for local spices. The island has many links to Africa. In fact most of the **(20)** _____ is of African origin. The climate is tropical and humid and there are beautiful beaches and coral reefs around the coast which makes it an attractive **(21)** _____for tourists. Although fishing and agriculture currently dominate the economy, the government is **(22)** _____tourism as a way of ensuring prosperity in the future.

Appendix 3 First Draft Expert Judges' Checklist

An Exploratory Study on the English Language Reading Test in the Foundation Programme in the Colleges of Technology, Oman

Expert Judge Checklist

Dear Colleague,

Thank you for your willingness to take part in this study. The amount of work and time commitment involved is greatly appreciated. Your time and effort will certainly make a valuable contribution the findings of this study.

The research is an exploration of English test in a second language given to students to assess their readiness for commencement of academic studies. What is required is your considered opinion on various aspects of test tasks compared with a selection of passages from first year academic text books. Each of the context features will be briefly defined to remind you of the particular being tested.

Your views will be treated with strict confidentiality and will not be revealed to any third party. Reporting will be in an anonymous form and your identity will not be disclosed. Your participation in this study is on a voluntary basis. You may withdraw from the study at any time before data analysis is carried out. Should you decide to withdraw, your right to do so will be fully respected and you will not be asked for a reason for your decision.

The questionnaire consists of **33 statements about second language assessment.**

You are kindly requested to answer all the questions in this part.

Completed questionnaires will be collected by me at the end of the session. Should you have any queries, please do not hesitate to ask me directly or contact me via my e-mail address: anwar-amar@hotmail.com

Once again, thank you for taking part in this study.

Sincerely,

Anwar Al-Ismaili
PhD researcher, School of Social and International Sciences
University of Bradford, United Kingdom

PART II: The Judgment Checklist

This checklist evaluates text/extract no:

What is required from you is your response to a set of statements using the Likert scale the ratings of which are explained in each statement. You will need to have your test paper/extracted text ready to answer the questions. **Circle the number** that most closely expresses your point of view. Remember that there is no correct or incorrect answer. Please answer all the questions. Remember to use a separate checklist for each text/extract and to fill in the text code in the given box

A. Linguistic Demands

Overall text purpose

Wiegle (2002, pp.8-10), based on the general model of writing discourse by (Vähäpässi, 1982), presents a model in which text types are divided into two categories; cognitive processing and main purpose. The main elements of overall purpose are based on

1. Identify the main elements of the overall text purpose:

	1	2	3	4	5
Options	Referential	Conative	Emotive	Poetic	Phatic
Examples	intended to inform	intended to persuade	intended to convey feelings or emotion	intended to entertain, delight, please	intended to keep in touch

Overall text purpose

The implication here is that test tasks relate to overall purpose.

2. The test tasks relate to the overall purpose:

	1	2	3	4	5
Options	Greatly relate	relate	fairly		
Examples	In referential texts the tasks must be confined to points of information that is explicit or perhaps implicit in the text.				

Writer-reader relationship

The intended reader or audience is of crucial importance in the creation of a text and its meaning. The intended audience determines the extent of content knowledge that the writer can assume that the reader already has.

3. Identify the intended audience/reader of the text that is targeted by the writer:

Options	**1** Audience addressed	**2** Audience invoked
Examples	is the intended reader	can be a fictitious reader, which often happens where the writer is writing for a rhetorical purpose

Discourse mode (Genre, rhetorical task, pattern of exposition, rhetorical organisation)

The reader's understanding of how texts are organized influences reading comprehension" In discourse mode, an argument or point of view or discussion is gradually built up in a logically or chronologically coherent way.

Genre:

1. Identify the most appropriate category.

Options	**1** text book	**2** magazine/newspaper article	**3** research/academic journal article	**4** report

Rhetorical task:

2. Identify the most appropriate category.

Options	**1** exposition	**2** argumentation/persuasion/evaluation	**3** historical biographical/autobiographical narrative

Pattern of exposition:

3. Identify the pattern(s) used in the text.

Options	**1** define	**2** describe	**3** elaborate	**4** illustrate	**5** compare /contrast	**6** classify	**7** cause/ effect	**8** problem /solution	**9** justify

Rhetorical organisation:

4. Does the text have an explicit organisational structure?

	1	2	3	4	5
Options	**Explicit**				**Not explicit**

Functional resources

Function is a term used to describe the illocutionary force of what is said. Examples of communicative functions might be where a speaker has to persuade, advise, describe, etc

5. Identify the most appropriate category.

Options	**1** Ideational	**2** Manipulative	**3** Heuristic	**4** Imaginative
Examples	Descriptions, classifications, explanations, and expressions of sorrow or anger	Requests, suggestions, commands, and warnings; rules, regulations and laws; greetings and leave-takings, compliments, insults, and apologies	for teaching and learning, problem solving, retention of information	Jokes, and use of figurative language and poetry

Grammatical resources

In the context of assessment, grammatical forms are important for inferring the exact meaning intended by the writer

Grammar:

6. The sentences in the text are:

Options	**1** mainly simple sentences	**2** a balance of simple and compound sentences	**3** mostly compound sentences	**4** a balance of compound and complex sentences	**5** mostly complex sentences

Cohesion

7. Throughout the text, are relations between the ideas explicitly marked through reference, conjunctions and connectors or are such relations not explicit?

Options	1	2	3	4	5
	(explicit)				(not explicit)

Nature of information (Text abstractness)

This refers to the degree to which a particular text is posed in predominantly abstract or concrete terms.

8. Is the text concrete or abstract?

Options	1	2	3	4	5
	(concrete)				(abstract)
Examples	The degrees of abstraction involved in the word 'chair' could be 'seat', then, 'furniture', then 'furnishings' and finally 'entity'.				

Content knowledge

The issue involved in content knowledge is to assess what influence the test taker's background knowledge may have on the relative difficulty of a specific test task.

9. Is the topic of the text of general interest or does it require subject specific knowledge on the part of the reader?

Options	1	2	3	4	5
	(general				(specific)

10. Is the topic of the text culture-neutral or is it loaded with specific cultural content?

Options	1	2	3	4	5
	(Culture neutral)				(cultural specific)

11. Is the topic of the text students language background-neutral or is it loaded for specific language background students?

Options	1	2	3	4	5
	(language background neutral)				(language background specific)

12. Is the topic of the text religion-neutral or is it loaded for specific with specific religion content?

Options	1	2	3	4	5
	(religion neutral)				(religion specific)

B. Task Setting

Response method (ONLY APPLIES TO TEST TASK PAPER) **DEFINITION and examples**					
13. Is there any evidence that the test response method format is likely to affect the test performance?					
Options	**1**	**2**	**3**	**4**	**5**
14. having a wide variety of types of tasks has been highly recommended as it provides multiple ways for learners to provide evidence of their strengths (The test tasks provides a variety of response methods					
Options	**1** Strongly agree	**2** Agree	**3** Undecided	**4** Disagree	**5** Strongly disagree

Weighting (ONLY APPLIES TO TEST TASK PAPER) Various tasks in a test can be assigned different maximum scores based on a belief of test designers that certain items are more important than others and should therefore carry more weight in scoring.					
15. Are any weighting for different test components adequately justifies?					
Options	**1** Adequately justifies	**2**	**3**	**4**	**5**
Examples					
16. The argument amounts to whether to apply weighting in the scoring or to score all items equally but for greater representation for more important items.					
Options May be only two options required	**1** Strongly agree	**2** Agree	**3** Undecided	**4** Disagree	**5** Strongly disagree
Example needed					

Knowledge of criteria (ONLY APPLIES TO TEST TASK PAPER)

If anything other than reading and comprehension is taken into account in scoring this must be stated in the criteria by which the work is assessed. In such a case writing skills are muddled up with reading and comprehension. (e.g. if spelling, punctuation, etc are being taken into account the candidate needs to know)

17. Are the criteria to be used in the marking of the test explicit for the candidates and the markers?

	1	2	3	4	5
Options	(explicit criteria)				(not explicit criteria)

Order of items (ONLY APPLIES TO TEST TASK PAPER)

Each sentence adds something to what has been established in earlier sentences so that the meaning gradually unfolds. Clearly, the order of items in a test in this situation is important.

18. Are the items and tasks in a test in a justifiable order?

	1	2	3	4	5
Options	(justifiable)				(not justifiable)

Channel of presentation (APPLIES TO TEST TASK PAPER and EXTRACTS)

Research suggests that comprehension is aided by information presented in more than one form. For example, text accompanied by a diagram, a picture or chart aids working memory.

19. Is the channel appropriate for the target situation requirements of the students being tested?

	1	2	3	4	5
Options	(appropriate)				(not appropriate)

Text length (APPLIES TO TEST TASK PAPER and EXTRACTS)

If it is intended to measure the ability of the testee to judge the relevance or irrelevance or distinguish between main points or subsidiary main details, according to Alderson (1996) a long text is needed for these operations to be truly realized (Khalifa and Weir, 2009, p. 99). Determining how long a text needs to be is not a straightforward task and most authors avoid being too prescriptive about the text length. The actual length will depend on the purpose of testing

20. Is the text length appropriate for the target situation requirements of the students being tested?

	1	2	3	4	5
Options	(appropriate)				(not appropriate)

Time constraint (ONLY APPLIES TO TEST TASK PAPER) might be helpful to include word count in test paper also include here 40 minutes for test

Authenticity demands that the cognitive skills involved in comprehension also need to be time constrained as in the real world of academic reading, time will likewise be limited.

21. Is the timing for each part of the test e.g. preparation and completion appropriate?

	1	2	3	4	5
Options	(appropriate)				(not appropriate)

Once again thank you for your time and effort

Appendix 4 Programmes offered by different academic departments for first year students

Engineering Department: Semester I

Course Code	Course Title	Required By:	Pre-Requisites	Passing Grade	Credit Hours	Theory Hours	Practical Hours
MATH1102	Pure Math	Department	FPMT0001	C-	3	4	0
ITAD1100	Advanced IT Skills	College	FPIT0001	D	3	0	6
CECE1100	Engineering Graphics	Department	None	C-	3	0	6
CHEM1100	Fundamentals Of Chemistry (Engineering)	Department	None	C-	3	2	2
ENTW1100	Technical Writing I	College	Have passed the LEE of Advanced Level in ELC	D	3	4	0

Semester II

Course Code	Course Title	Required By:	Pre-Requisites	Passing Grade	Credit Hours	Theory Hours	Practical Hours
ENTW1200	Technical Writing II	College	ENTW1100	D	3	4	0
MATH1200	Calculus I	Department	Math1102	C-	3	4	0
EEPW1240	Engineering Workshop	Department	None	C-	3	0	6
PHYS1100	Physics I (Engineering)	Department	MATH1102	C-	3	2	2
EECP1290	Computer Programming for Engineering	Department	ITSE1100	C-	3	0	6

Semester III

Course Code	Course Title	Required By:	Pre-Requisites	Passing Grade	Credit Hours	Theory Hours	Practical Hours
PHYS1211	Physics II	Department	PHYS1100	C-	3	2	2
BACO1212	Job Search Techniques	College	None	D	3	4	0

Business Studies: Semester I

Course Code	Course Title	Required By:	Pre-Requisites	Passing Grade	Credit Hours	Theory Hours	Practical Hours
ENTW 1100	Technical Writing I	College	None	D	3	2	2
BAMG 1100	Introduction To Business	Department	None	C-	3	2	2
MATH 1103	Applied Mathematics	Department	FPMT0001	Non Credit Course		2	2
ITAD 1100	Advanced IT Skills	College	None	D	3	0	6
BAAC 1102	Principles Of Accounting	Department	None	C-	3	1	4
Total Semester Credit Hrs:					**12**	**7**	**16**

Semester II

Course Code	Course Title	Required By:	Pre-Requisites	Passing Grade	Credit Hours	Theory Hours	Practical Hours
ENTW 1200	Technical Writing II	College	ENTW 1100	D	3	2	2
BAMK 1205	Principles Of Marketing	Department	BAMG 1100	C-	3	2	2
BAEC 1203	Principles Of Microeconomics	Department	None	C-	3	2	2
BAST 1206	Managerial Statistics	Department	BAMA 1101	C-	3	2	2
BAMG 1207	Principles of Management	Department	MATH 1103	C-	3	2	2
Total Semester Credit Hrs:					**15**	**10**	**10**

Semester II

Course Code	Course Title	Required By:	Pre-Requisites	Passing Grade	Credit Hours	Theory Hours	Practical Hours
ENGL 1208	Business Communication	Department	ENTW 1200	C	3	1	4
BACO 1212	Job Search Tecniques	College	None	D	3	1	4
Total Semester Credit Hrs:					**6**	**2**	**8**

IT department Semester I

Course Code	Course Title	Required By	Pre-Requisites	Passing Grade	Credit Hours	Theory Hours	Practical Hours
MATH1102	Pure Math	College	FPMT0001	D	3	4	0
ITSE1100	Information System Multimedia	College	GFP Computing (FP Level4)	D	3	0	6
ITNT1103	Computer Hardware	Specialization	None	C	3	1	4
ITAD1100	Advanced IT Skills	College	FPIT01	D	3	0	4
ENTW1100	Technical Writing I	College	The student should have passed the LEE of Advanced Level in ELC (FP Level4)	D	3	4	0
ITDB1102	Introduction To Database	Specialization	None	C	3	1	4
Total Semester Credit Hrs:					**18**	**10**	**18**

Semester II

Course Code	Course Title	Required By	Pre-Requisites	Passing Grade	Credit Hours	Theory Hours	Practical Hours
ITSE1101	Introduction To Programming C	Specialization	None	C	3	1	4
MATH1201	Managerial Statistics	Department	Math1101	C-	3	2	2
ENTW1200	Technical Writing II	College	ENTW1100	D	3	4	0
ITDB1204	Applied Database	Specialization	ITDB1102	C	3	1	4
BACO1212	Job Search Tecniques	College	None	D	3	4	0
ITNT1204	Introduction To Networking	Specialization	ITNT1103	C	3	1	4
Total Semester Credit Hrs:					**18**	**13**	**14**

Semester III

Course Code	Course Title	Required By	Pre-Requisites	Passing Grade	Credit Hours	Theory Hours	Practical Hours
ITSE1202	Introduction To Operating System	Specialization	None	C	3	1	4
ITSE1203	Introduction To Web Technology	Specialization	None	C	3	0	6
Total Semester Credit Hrs:					**6**	**1**	**10**

Adopted from Ibra College of Technology (2014a, 2014b, 2014c)

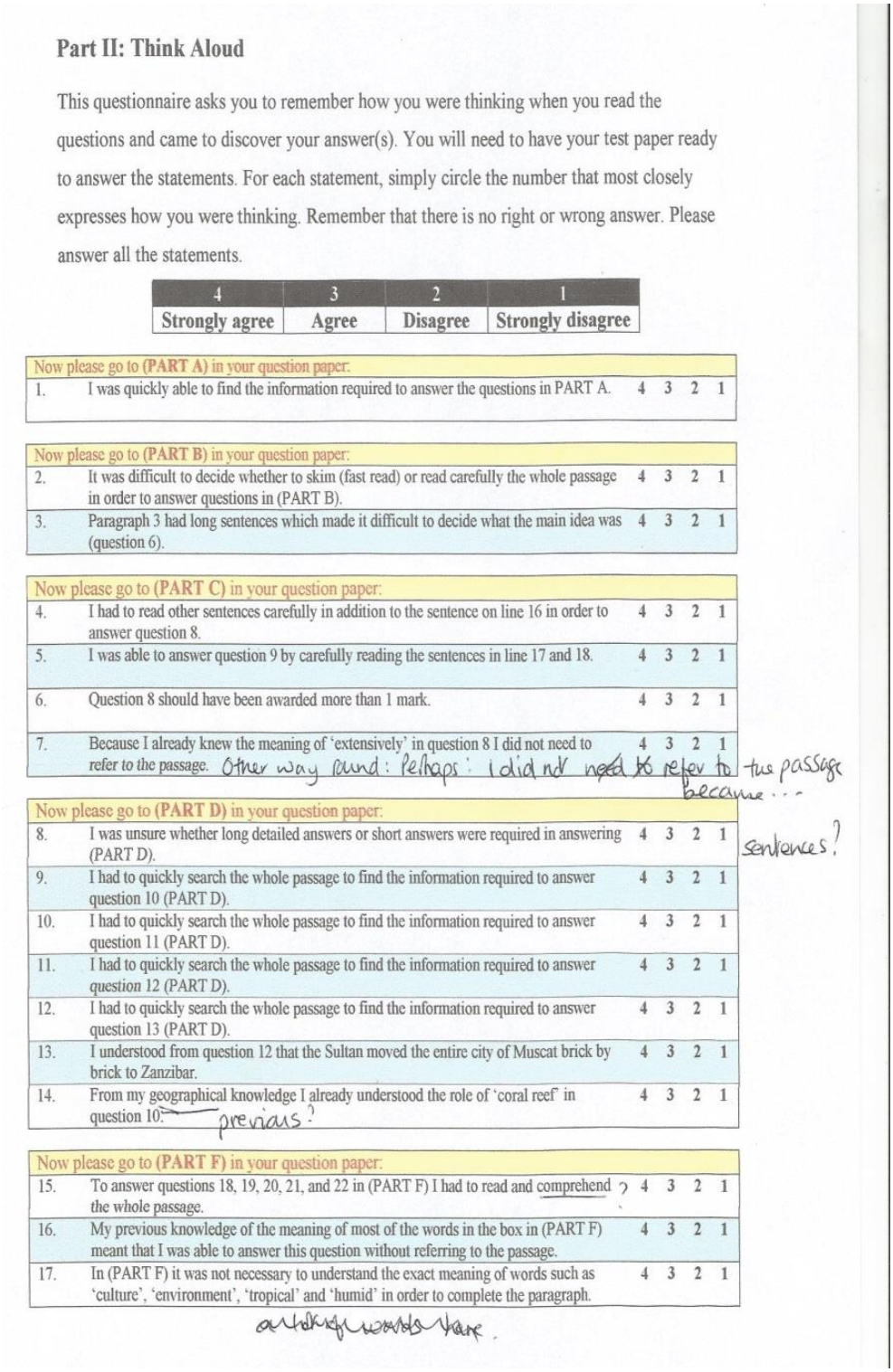
Appendix 5 Instrument Piloting: Questionnaire evaluation checklist

1. How long did it take you to complete?
2. Were the instructions clear?
3. Were any of the questions unclear or ambiguous? If so, will you say which and why?
4. Did you object to answering any of the questions?
5. In your opinion, has any major issue been omitted? Please specify.
6. Was the layout of the questionnaire clear/attractive?
7. Any comments?

Adapted and reproduced in a form of checklist from (Bell, 2010, p.15)

Appendix 6 Draft II student questionnaire used for trialling

Piloting Stage II – Student Questionnaire Editing



Part II: Think Aloud

This questionnaire asks you to remember how you were thinking when you read the questions and came to discover your answer(s). You will need to have your test paper ready to answer the statements. For each statement, simply circle the number that most closely expresses how you were thinking. Remember that there is no right or wrong answer. Please answer all the statements.

4	3	2	1
Strongly agree	Agree	Disagree	Strongly disagree

Now please go to **(PART A)** in your question paper:					
1.	I was quickly able to find the information required to answer the questions in PART A.	4	3	2	1
Now please go to **(PART B)** in your question paper:					
2.	It was difficult to decide whether to skim (fast read) or read carefully the whole passage in order to answer questions in (PART B).	4	3	2	1
3.	Paragraph 3 had long sentences which made it difficult to decide what the main idea was (question 6).	4	3	2	1
Now please go to **(PART C)** in your question paper:					
4.	I had to read other sentences carefully in addition to the sentence on line 16 in order to answer question 8.	4	3	2	1
5.	I was able to answer question 9 by carefully reading the sentences in line 17 and 18.	4	3	2	1
6.	Question 8 should have been awarded more than 1 mark.	4	3	2	1
7.	Because I already knew the meaning of 'extensively' in question 8 I did not need to refer to the passage. Other way round: Perhaps: I did not need to refer to the passage because...	4	3	2	1
Now please go to **(PART D)** in your question paper:					
8.	I was unsure whether long detailed answers or short answers were required in answering (PART D).	4	3	2	1
9.	I had to quickly search the whole passage to find the information required to answer question 10 (PART D).	4	3	2	1
10.	I had to quickly search the whole passage to find the information required to answer question 11 (PART D).	4	3	2	1
11.	I had to quickly search the whole passage to find the information required to answer question 12 (PART D).	4	3	2	1
12.	I had to quickly search the whole passage to find the information required to answer question 13 (PART D).	4	3	2	1
13.	I understood from question 12 that the Sultan moved the entire city of Muscat brick by brick to Zanzibar.	4	3	2	1
14.	From my geographical knowledge I already understood the role of 'coral reef' in question 10. previous?	4	3	2	1
Now please go to **(PART F)** in your question paper:					
15.	To answer questions 18, 19, 20, 21, and 22 in (PART F) I had to read and comprehend the whole passage.	4	3	2	1
16.	My previous knowledge of the meaning of most of the words in the box in (PART F) meant that I was able to answer this question without referring to the passage.	4	3	2	1
17.	In (PART F) it was not necessary to understand the exact meaning of words such as 'culture', 'environment', 'tropical' and 'humid' in order to complete the paragraph.	4	3	2	1

sentences?

Now please go to (the passage):					
18.	The use of 'At the moment', 'but' in lines 33 and 34 helped me to understand the present tourist situation in Zanzibar and future projections.	4	3	2	1
19.	In lines 9 and 10, the structure 'though….', 'now' helped me to understand current developments of conservation of trees.	4	3	2	1
20.	In line 7 the sentence 'the climate is tropical and humid', required the sentence that followed it for a more in- depth understanding.	4	3	2	1
21.	The 'colourful history' of Zanzibar in line 27 is due to the amount of sunshine the island receives.	4	3	2	1
22.	The sentence in line 30 would be more meaningful if it read 'the people in Zanzibar also manufacture clove oil and woven goods'	4	3	2	1
23.	It was easy to understand what was referred to by 'it' in line 7	4	3	2	1
24.	Although paragraph 1 had only two short sentences it was difficult to decide which of the three options best expressed the main idea.	4	3	2	1
25.	It was necessary to refer to the passage to work out the meaning of "prosperous" in line 17.	4	3	2	1

Miscellaneous issues:					
26.	I found the test instruction easy to understand	4	3	2	1
27.	I agree with the number of marks given for each task	4	3	2	1
28.	I found the variety of question types (e.g. true/false, short answer, gap filling, etc.) helpful in allowing me to show my skills.	4	3	2	1
29.	It was helpful to know how many marks were allocated to each item.	4	3	2	1
30.	The knowledge of the number of marks given to each item affected my time planning and execution	4	3	2	1
31.	Before starting to answer the questions it was important to plan how much time should be spent on each item.	4	3	2	1
32.	It was important to answer the questions in the order they were presented.	4	3	2	1
33.	A time line of main events would have aided and supported my passage comprehension.	4	3	2	1
34.	It was not necessary to read every word of the passage in order to understand its meaning.	4	3	2	1
35.	The length of the passage made it difficult to read and understand in the time provided.	4	3	2	1
36.	I looked at the questions first before deciding whether to read the passage carefully or quickly.	4	3	2	1
37.	40 minutes was sufficient time to answer all the questions.	4	3	2	1
38.	Recognising the overall purpose of the passage helped my understanding.	4	3	2	1
39.	This article is probably taken from a secondary school history passage book.	4	3	2	1
40.	The organisation of the five paragraphs in the passage helped my understanding and comprehension.	4	3	2	1
41.	This passage is suitable for Omani students.	4	3	2	1
42.	The passage presents the historical development of Zanzibar in an easy- to- understand way.	4	3	2	1
43.	Most of the words on the passages were concrete, ~~meaning~~ factual ~~straight~~ words and easy to understand, e.g. car, apple rather than abstract words e.g. economy	4	3	2	1
44.	The article was mostly familiar to me due to my knowledge of the history of the connections between Zanzibar and Oman.	4	3	2	1
45.	Even though there were no headings, it was still easy to understand the passage.	4	3	2	1

Thank you for your participation

Part II: Think Aloud

This questionnaire asks you to remember how you were thinking when you read the questions and came to discover your answer(s). You will need to have your test paper ready to answer the statements. For each statement, simply circle the number that most closely expresses how you were thinking. Remember that there is no right or wrong answer. Please answer all the statements.

4	3	2	1
Strongly agree	**Agree**	**Disagree**	**Strongly disagree**

Now please go to **(PART A)** in your question paper:					
1.	I was quickly able to find the information required to answer the questions in PART A.	4	3	2	1
Now please go to **(PART B)** in your question paper:					
2.	It was difficult to decide whether to skim (fast read) or read carefully the whole passage in order to answer questions in (PART B). Same passage, or read the whole passage carefully	4	3	2	1
3.	Paragraph 3 had long sentences which made it difficult to decide what the main idea was (question 6).	4	3	2	1
Now please go to **(PART C)** in your question paper:					
4.	I had to read other sentences carefully in addition to the sentence on line 16 in order to answer question 8.	4	3	2	1
5.	I was able to answer question 9 by carefully reading the sentences in line 17 and 18.	4	3	2	1
6.	Question 8 should have been awarded more than 1 mark.	4	3	2	1
7.	I did not need to refer to the passage because I already knew the meaning of '*extensively*' *in question* 8.	4	3	2	1
Now please go to **(PART D)** in your question paper:					
8.	I was unsure whether long detailed answers or short answers were required in answering (PART D).	4	3	2	1
9.	I had to quickly search the whole passage to find the information required to answer question 10 (PART D).	4	3	2	1
10.	I had to quickly search the whole passage to find the information required to answer question 11 (PART D).	4	3	2	1
11.	I had to quickly search the whole passage to find the information required to answer question 12 (PART D).	4	3	2	1
12.	I had to quickly search the whole passage to find the information required to answer question 13 (PART D).	4	3	2	1
13.	I understood from question 12 that the Sultan moved the entire city of Muscat brick by brick to Zanzibar.	4	3	2	1
14.	From my geographical knowledge I already understood the role of 'coral reef' in question 10.	4	3	2	1
Now please go to **(PART F)** in your question paper:					
15.	To answer questions 18, 19, 20, 21, and 22 in (PART F) I had to read and understand the whole passage.	4	3	2	1
16.	My previous knowledge of the meaning of most of the words in the box in (PART F) meant that I was able to answer this question without referring to the passage.	4	3	2	1
17.	In (PART F) it was not necessary to understand the exact meaning of words such as 'culture', 'environment', 'tropical' and 'humid' in order to complete the paragraph.	4	3	2	1

Now please go to **(the passage)**:					
18.	The use of 'At the moment', 'but' in lines 33 and 34 helped me to understand the present tourist situation in Zanzibar and future projections.	4	3	2	1
19.	In lines 9 and 10, the structure 'though....', 'now' helped me to understand current developments of conservation of trees.	4	3	2	1
20.	In line 7 the sentence 'the climate is tropical and humid', required the sentence that followed it for a more in- depth understanding.	4	3	2	1
21.	The 'colourful history' of Zanzibar in line 27 is due to the amount of sunshine the island receives.	4	3	2	1
22.	The sentence in line 30 would be more meaningful if it read 'the people in Zanzibar also manufacture clove oil and woven goods'	4	3	2	1
23.	It was easy to understand what was referred to by 'it' in line 7	4	3	2	1
24.	Although paragraph 1 had only two short sentences it was difficult to decide which of the three options best expressed the main idea.	4	3	2	1
25.	It was necessary to refer to the passage to work out the meaning of "prosperous" in line 17.	4	3	2	1

Miscellaneous issues:					
26.	I found the test instruction easy to understand.	4	3	2	1
27.	I agree with the number of marks given for each task,	4	3	2	1
28.	I found the variety of question types (e.g. true/false, short answer, gap filling, etc.) helpful in allowing me to show my skills.	4	3	2	1
29.	It was helpful to know how many marks were allocated to each item.	4	3	2	1
30.	The knowledge of the number of marks given to each item affected my time planning and execution	4	3	2	1
31.	Before starting to answer the questions it was important to plan how much time should be spent on each item.	4	3	2	1
32.	It was important to answer the questions in the order they were presented.	4	3	2	1
33.	A time line of main events would have aided and supported my passage comprehension. (my comp. of the passage)	4	3	2	1
34.	It was not necessary to read every word of the passage in order to understand its meaning.	4	3	2	1
35.	The length of the passage made it difficult to read and understand in the time provided.	4	3	2	1
36.	I looked at the questions first before deciding whether to read the passage carefully or quickly.	4	3	2	1
37.	40 minutes was sufficient time to answer all the questions. (a sufficient amount of time)	4	3	2	1
38.	Recognising the overall purpose of the passage helped my understanding.	4	3	2	1
39.	This article is probably taken from a secondary school history passage book.	4	3	2	1
40.	The organisation of the five paragraphs in the passage helped my understanding and comprehension.	4	3	2	1
41.	This passage is suitable for Omani students.	4	3	2	1
42.	The passage presents the historical development of Zanzibar in an easy- to- understand way.	4	3	2	1
43.	Most of the words in the passages were concrete, meaning factual straight words and easy to understand, e.g. car, apple rather than abstract words e.g. economy (straightforward)	4	3	2	1
44.	The article was mostly familiar to me due to my knowledge of the history of the connections between Zanzibar and Oman.	4	3	2	1
45.	Even though there were no headings, it was still easy to understand the passage.	4	3	2	1

Thank you for your participation

Instrument Piloting

Questionnaire Evaluation Checklist

1. How long did it take you to complete? 25 minutes
2. Were the instructions clear? Mostly, yes
3. Were any of the questions unclear or ambiguous? If so, will you say which and why? Q 43 was difficult using words "concrete" and "abstract" were difficult
4. Did you object to answering any of the questions? No
5. In your opinion, has any major issue been omitted? Please specify. No
6. Was the layout of the questionnaire clear/attractive? Yes
7. Any comments? Maybe too many questions - otherwise it seems good

Thank you for your time and effort

Instrument Piloting

Really thorough and well thought out questionnaire good luck with pilot and research!

Gwyn

Questionnaire Evaluation Checklist

1. How long did it take you to complete? Very quick and simple
2. Were the instructions clear? Yes
3. Were any of the questions unclear or ambiguous? If so, will you say which and why? use of commas, grammar inconsistent but very understandable
4. Did you object to answering any of the questions? no
5. In your opinion, has any major issue been omitted? Please specify.
6. Was the layout of the questionnaire clear/attractive? Yes
7. Any comments? ① Cognition & Second Language Instruction Robinson (ed) 2001 *② Cognitive Processing in Second language Aqisition Inside the learner's mind Pütz & Sicola (eds) 2010 (on google Books)

Thank you for your time and effort

Appendix 7 Final version of the student questionnaire used for VPA (Arabic and English)

Part I: Background Information معلومات شخصية

Student No.

1. **Gender (Please circle the appropriate number below):**

1	2
Male ذكر	Female أنثى

2. **Subject specialisation (Please circle the appropriate number below):**

1	2	3	4	5	6	7
Engineering هندسه	IT نظم معلوما ت	Business Studies دراسات تجاريه	Applied Sciences العلوم التطبيقية	Pharmacy الصيدلة	Fashion Design تصميم الأزياء	Photography التصوير الفوتوغرافي

Part II: Think Aloud فكر بصوت عالي

This questionnaire asks you to remember how you were thinking when you read the questions and came to discover your answer(s). You will need to have your test paper ready to answer the statements. For each statement, simply circle the number that most closely expresses how you were thinking. Remember that there is no right or wrong answer. Please answer all the statements.

من المتأمل منك في هذا الإستبيان أن تتذكر ما تفكر به و أنت تقرأ أسئلة الإمتحان و كيف توصلت إلى الإجابات. الرجاء الرجوع لورقة الإمتحان للتمكن جيدا من الإجابة على الإستبيان. لكل عبارة، الرجاء عمل دائرة حول رقم الإجابة المناسبة و التي تعبر عن ما كنت تفكر به. تذكر دائما بأنه لا توجد هناك عبارة صحيحة أو خطأ. الرجاء الإجابة على كل العبارات.

4	3	2	1
Strongly agree	**Agree**	**Disagree**	**Strongly disagree**

Now please go to **(PART A)** in your question paper: **الآن الرجاء الرجوع ل (PART A) في ورقة الإمتحان للتمكن من إجابة العبارات التالية:**					
1.	I was quickly able to find the information required to answer the questions in PART A. **كان بإستطاعتي بسرعة إيجاد المعلومات المطلوبة للإجابة على الأسئلة في (الجزأ A)**	**4**	**3**	**2**	**1**

Now please go to **(PART B)** in your question paper: **الآن الرجاء الرجوع ل (PART B) في ورقة الإمتحان للتمكن من إجابة العبارات التالية:**					
2.	I found it difficult to decide whether to skim (fast read) or read carefully the whole passage in order to answer questions. **صادفت صعوبة لكي أقرر أن أقرأ النص كاملا بسرعة أو أقرأ بتمهل للإجابة على الأسئلة.**	**4**	**3**	**2**	**1**
3.	I found it difficult to decide what was the main idea in Paragraph 3 because the sentences were too long (question 6). **صادفت صعوبة لكي أقرر ما هي الفكرة الرئيسية في الفقرة 3 لأن الجمل كانت طويلة جدا.**	**4**	**3**	**2**	**1**

Now please go to **(PART C)** in your question paper: **الآن الرجاء الرجوع ل (PART C) في ورقة الإمتحان للتمكن من إجابة العبارات التالية:**					
4.	I had to read other sentences carefully in addition to the sentence on line 16 in order to answer question 8. **إضطررت لقراءة جمل أخرى بتمهل إضافة الى الجملة في السطر 16 لكي أجاوب على السؤال 8.**	**4**	**3**	**2**	**1**
5.	I was able to answer question 9 by carefully reading the sentences in line 17 and 18. **إستطعت الإجابة على السؤال 9 و ذلك عن طريق قراءة الجمل في السطور 17 و 18 بتمهل.**	**4**	**3**	**2**	**1**
6.	I did not need to refer to the passage to answer question 8 because I already knew the meaning of 'extensively'. **لم أحتاج للرجوع إلى النص لكي أجاوب السؤال 8 و ذلك بسبب معرفتي مسبقا لمعنى كلمة 'extensively'.**	**4**	**3**	**2**	**1**

Now please go to **(PART D)** in your question paper: **الآن الرجاء الرجوع ل (PART D) في ورقة الإمتحان للتمكن من إجابة العبارات التالية:**					
7.	I was unsure whether long detailed answers or short answers were required in answering Part D. **لم أكن متأكدا مما إذا كانت الإجابات المطولة كثيرة التفاصيل أو الإجابات القصيرة مطلوبة للإجابة على السؤال.**	**4**	**3**	**2**	**1**
8.	I had to quickly search the whole passage to find the information required to answer question 10. **كنت مضطرا للبحث بسرعة في النصه كله لكي أجد المعلومات المطلوبة للإجابة عن السؤال 10.**	**4**	**3**	**2**	**1**
9.	I had to quickly search the whole passage to find the information required to answer question 11. **كنت مضطرا للبحث بسرعة في النصه كله لكي أجد المعلومات المطلوبة للإجابة عن السؤال 11.**	**4**	**3**	**2**	**1**
10.	I had to quickly search the whole passage to find the information required to answer question 12. **كنت مضطرا للبحث بسرعة في النصه كله لكي أجد المعلومات المطلوبة للإجابة عن السؤال 12.**	**4**	**3**	**2**	**1**
11.	I had to quickly search the whole passage to find the information required to answer question 13. **كنت مضطرا للبحث بسرعة في النصه كله لكي أجد المعلومات المطلوبة للإجابة عن السؤال 13.**	**4**	**3**	**2**	**1**
12.	I understood from question 12 that the Sultan moved the entire city of Muscat brick by brick to Zanzibar. **فهمت من السؤال 12 بأن السلطان نقل مدينته كاملة بمبانيها من مسقط إلى زنجبار.**	**4**	**3**	**2**	**1**

Now please go to **(PART F)** in your question paper: **الآن الرجاء الرجوع ل (PART F) في ورقة الإمتحان للتمكن من إجابة العبارات التالية:**					
13.	To answer questions 18, 19, 20, 21, and 22, I had to read and understand the whole passage. **لكي أتمكن من الإجابة على الأسئلة 18، 19، 20، 21، و 22، إضطررت لقراءة و فهم النص كاملا.**	**4**	**3**	**2**	**1**
14.	I did not need to know the exact meaning of the words such as 'culture', 'environment', 'tropical' and 'humid' in order to complete the paragraph. **لم أحتاج لمعرفة معاني الكلمات الآتية بالتحديد ('culture', 'environment', 'tropical' and 'humid') لكي أكمل الفقرة بشكل صحيح.**	**4**	**3**	**2**	**1**

Now please go to **(the passage)**: **الآن الرجاء الرجوع للنص للتمكن من إجابة العبارات التالية:**					
15.	I needed to read beyond the sentence in line 7 the sentence in order to understand the meaning of 'the climate is tropical and humid'. **إحتجت للقراءة أبعد من الجملة في السطر 7 لكي أجاوب على معنى 'the climate is tropical and humid'.**	**4**	**3**	**2**	**1**
16.	I think that the history of Zanzibar is described as 'colourful' in line 27 due to the amount of sunshine that Zanzibar receives. **أعتقد أن تاريخ زنجبار بوصفه بكلمة 'colourful' في السطر 27 يرجع إلى شروق الشمس الدائم في زنجبار.**	**4**	**3**	**2**	**1**
17.	I could easily understand 'it' referred to in line 7. **كان بإستطاعتي بكل سهولة و يسر أن أفهم ما تعود إليه 'it' في السطر 7.**	**4**	**3**	**2**	**1**
18.	I found it easy to decide which of the three options best express the main idea of Paragraph 1 because the sentences were short. **كان سهلا أن أقرر أي من الخيارات الثلاثة يعبر عن الفكرة الرئيسية في الفقرة 1 لأن الجمل كانت قصيرة.**	**4**	**3**	**2**	**1**

Miscellaneous issues: **قضايا مختلفة:**					
19.	I found the test instructions easy to understand.**وجدت تعليمات الإمتحان سهلة الفهم**	**4**	**3**	**2**	**1**
20.	I think the number of marks given for each question was appropriate. **أعتقد بأن الدرجات المخصصة لكل سؤال مناسبة.**	**4**	**3**	**2**	**1**
21.	I found the variety of question types (e.g. true/false, short answer, gap filling, etc.) helpful in allowing me to show my skills. **أعتقد بأن التنوع في الأسئلة (صح\خطأ، إجابات قصيرة، إملأ الفراغات) قد ساعدتني لإبراز مهاراتي المختلفة.**	**4**	**3**	**2**	**1**
22.	It was helpful to know how many marks were allocated to each item. **معرفة الدرجات المخصصة لكل سؤال ساعدني كثيرا.**	**4**	**3**	**2**	**1**

23.	I thought it was important to answer the questions in the order they were presented. **أعتقدت أنه من المهم الإجابة على الأسئلة على النحو و الترتيب التي جاءت عليه كما هو في ورقة الأسئلة.**	**4**	**3**	**2**	**1**
24.	A time line of main events would have aided and supported my passage comprehension. **لو تم إضافة خط أحداث أو جدول أو رسم بياني لكان بالإمكان أن يساعدني لفهم النص.**	**4**	**3**	**2**	**1**
25.	I did not need to read every word of the passage in order to understand its meaning. **لم أحتاج إلى قراءة كل كلمة في النص لفهم معانيها.**	**4**	**3**	**2**	**1**
26.	I looked at the questions first before deciding whether to read the passage carefully or quickly. **إضطررت للنظر و الرجوع إلى الأسئلة قبل أن أقرر لقراءة النص بتمهل أو بسرعة.**	**4**	**3**	**2**	**1**
27.	40 minutes was sufficient time to answer all the questions. **40 دقيقة كانت كافية للإجابة على الأسئلة.**	**4**	**3**	**2**	**1**
28.	Recognising the overall purpose of the passage helped my understanding. **التعرف على الغاية الكلية للنص ساعدني على الفهم.**	**4**	**3**	**2**	**1**
29.	I think the way the paragraphs are arranged helped me to understand the passage. **أعتقد بأن الطريقة التي تم ترتيب الفقرات بها ساعدتني على فهم النص.**	**4**	**3**	**2**	**1**
30.	This passage is suitable for Omani students.**.النص كان مناسبا للطلبة العمانيين**	**4**	**3**	**2**	**1**
31.	The passage presents the historical development of Zanzibar in an easy - to-understand way. **يعرض النص الأحداث و التطورات التاريخية في زجبار بشكل سلس و سهل للفهم.**	**4**	**3**	**2**	**1**
32.	I found most of the words on the passages were factual e.g. car, apple. **أجد معظم الكلمات في النص كلمات حقيقية و واقعية (جماد مثل car, apple).**	**4**	**3**	**2**	**1**
33.	The article was mostly familiar to me due to my knowledge of the history of the connections between Zanzibar and Oman. **كان النص مألوفا لدي بسبب معرفتي بتاريخ زنجبار و إرتباط زنجبار بعمان.**	**4**	**3**	**2**	**1**
34.	I was easily able to understand the passage even though there were no headings. **كنت قادرا على فهم النص رغم أنه لم توجد أي عناوين أو عناوين فرعية لكل فقرة.**	**4**	**3**	**2**	**1**

Thank you for your participation

Appendix 8 The expert judges' checklist

An Exploratory Study on the English Language Reading Test in the Foundation Programme in the Colleges of Technology, Oman

Expert Judge Checklist

Dear Colleague,

Thank you for your willingness to take part in this study. The amount of work and time commitment involved is greatly appreciated. Your time and effort will certainly make a valuable contribution the findings of this study.

The research is an exploration of the English as a second language test given to students to assess their readiness for commencement of academic studies. What is required is your considered opinion on various aspects of test tasks compared with a selection of passages from first year academic text books. Each of the context features will be briefly defined to remind you of the particular feature being tested.

Your views will be treated with strict confidentiality and will not be revealed to any third party. Reporting will be in an anonymous form and your identity will not be disclosed. Your participation in this study is on a voluntary basis. You may withdraw from the study at any time before data analysis is carried out. Should you decide to withdraw, your right to do so will be fully respected and you will not be asked for a reason for your decision.

The checklist consists of 22 statements about second language assessment.

You are kindly requested to answer all the questions in this part. When you have completed the checklist, I would be grateful if you could return them to me at your convenience. I will need to receive your responses no later than two weeks time (date). However, your contribution to this research is greatly valued and I hope it will be possible for you to meet this deadline. Should you have any queries, please do not hesitate to ask me directly or contact me via my e-mail address: anwar-amar@hotmail.com

Once again, thank you for taking part in this study.

Sincerely,

Anwar Al-Ismaili

PhD researcher, School of Social and International Sciences

University of Bradford, United Kingdom

Judgment Checklist – TEST TASK VERSION

This checklist evaluates text/extract no:

What is required from you is your response to a set of statements using the Likert scale, the ratings of which are explained in each statement. You will need to have your test paper/text extract ready to complete the checklist. **Circle the number** that most closely expresses your point of view. Remember that there is no correct or incorrect answer. Please respond to each item.

A. Linguistic Demands

Overall text purpose

Since purpose and task both relate to the choice of text, a consideration of text type and topic is crucial to content validity.

1. The category that best describes the overall text purpose is…

Options	**1** Referential	**2** Conative	**3** Emotive	**4** Poetic	**5** Phatic
Examples	intended to inform	intended to persuade	intended to convey feelings or emotion	intended to entertain, delight, please	intended to keep in touch

Writer-reader relationship

The intended audience determines the extent of content knowledge that the writer can assume that the reader already has.

2. Identify the intended audience/reader of the text that is targeted by the writer.

Options	**1** Audience addressed	**2** Audience invoked
Examples	is the intended reader	can be a fictitious reader, which often happens where the writer is writing for a rhetorical purpose

Discourse mode (Genre, rhetorical task, pattern of exposition, rhetorical organisation)
The reader's understanding of how texts are organized influences reading comprehension". In discourse mode, an argument or point of view or discussion is gradually built up.

Genre:

3. Identify the most appropriate category for the text.

	1	2	3	4
Options	text book	magazine/newspaper article	research/academic journal article	report

Rhetorical task:

4. Identify the most appropriate category for the text.

	1	2	3
Options	exposition	argumentation/persuasion/ evaluation	historical biographical/ autobiographical narrative

Pattern of exposition:

5. Identify the pattern(s) used in the text *(You may decide that more than one option applies).*

	1	2	3	4	5	6	7	8	9
Options	define	describe	elaborate	illustrate	compare /contrast	classify	cause/ effect	problem /solution	justify

Rhetorical organisation:

6. **The organisational structure of the text is…**

	1	2	3	4	5
Options	Explicit				Not explicit

Functional resources
The illocutionary force of what is said. Examples include situations where a speaker has to persuade, advise, describe, etc

7. Identify the most appropriate category for the text.

Options	**1** Ideational	**2** Manipulative	**3** Heuristic	**4** Imaginative
Examples	Descriptions, classifications, explanations, and expressions of sorrow or anger	Requests, suggestions, commands, and warnings; rules, regulations and laws; greetings and leave-takings, compliments, insults, and apologies	for teaching and learning, problem solving, retention of information	Jokes, and use of figurative language and poetry

Grammatical resources (grammar, cohesion)

Grammatical forms are important for inferring the exact meaning intended by the writer.

Grammar:

8. The sentences in the text are:

Options	**1** mainly simple sentences	**2** a balance of simple and compound sentences	**3** mostly compound sentences	**4** a balance of compound and complex sentences	**5** mostly complex sentences

Cohesion

9. Throughout the text, are relations between the ideas explicitly marked through reference, conjunctions and connectors or are such relations not explicit?

Options	**1** (explicit)	**2**	**3**	**4**	**5** (not explicit)

Nature of information (Text abstractness)

The degree to which a particular text is posed in predominantly abstract or concrete terms.

10. Is the text concrete or abstract?

Options	**1** (concrete)	**2**	**3**	**4**	**5** (abstract)
Examples	The degrees of abstraction involved in the word ‘chair’ could be ‘seat’, then, ‘furniture’, then ‘furnishings’ and finally ‘entity’.				

Content knowledge (General, cultural, language, religion)

What influence the test taker’s background knowledge may have on the relative difficulty of a specific test task.

11. Is the topic of the text of general interest or does it require subject specific knowledge on the part of the reader?

Options	**1** (general	**2**	**3**	**4**	**5** (specific)

Cultural background:

12. Is the topic of the text culture-neutral or is it loaded with specific cultural content?

Options	1 (Culture neutral)	2	3	4	5 (cultural specific)

Language background:

13. The text is significantly easier to understand for readers from a specific first language background

Options	1 Strongly agree	2	3	4	5 Strongly disagree

Religion knowledge:

14. Is the topic of the text religion-neutral or is it loaded with specific religious content?

Options	1 (religion neutral)	2	3	4	5 (religion specific)

Thank you for taking time to complete

(A. Linguistic Demands)

⬇⬇⬇

Please turn this page over and continue with questions on (B. Task Setting)

B. Task Setting

Response method

The mode of the test takers response e.g Yes/No, True/False, Gap Fill etc.

15. The test response method format is likely to affect the test performance?

Options	**1** Strongly agree	**2** Agree	**3** Undecided	**4** Disagree	**5** Strongly disagree

16. The test tasks provide a variety of response methods

Options	**1** Strongly agree	**2** Agree	**3** Undecided	**4** Disagree	**5** Strongly disagree

Weighting

Various tasks in a test can be assigned different maximum scores based on a belief of test designers that certain items are more important than others and should therefore carry more weight in scoring.

17. In general, the weighting for different test components are…

Options	**1** (Justified)	**2**	**3**	**4**	**5** (Not justified)

Knowledge of criteria

If anything other than reading and comprehension is taken into account in scoring this must be stated in the criteria by which the work is assessed. (e.g. if spelling, punctuation, etc are being taken into account the candidate needs to know)

18. The criteria to be used in the marking of the test for the candidates and the markers are…

Options	**1** (explicit criteria)	**2**	**3**	**4**	**5** (not explicit criteria)

Order of items

Each sentence adds something to what has been established in earlier sentences so that the meaning gradually unfolds.

19. The items and tasks in the test are presented in a justifiable order.

	1	2	3	4	5
Options	(justifiable)				(not justifiable)

Channel of presentation

Comprehension is aided by information presented in more than one form. For example, text accompanied by a diagram, a picture or chart aids working memory.

20. The channel for the target situation requirements of the students being tested is…

	1	2	3	4	5
Options	(appropriate)				(not appropriate)

Text length

Of sufficient length and size to provide ample opportunities for test takers to demonstrate their reading and comprehension. The actual length will depend on the purpose of testing.

21. The text length for the target situation requirements of the students being tested is…

	1	2	3	4	5
Options	(appropriate)				(not appropriate)

Time constraint

Authenticity demands that the cognitive skills involved in comprehension also need to be time constrained in the real world of academic reading, time will likewise be limited.

22. The test time of 40 minutes for the test (e.g. preparation and completion) is...

Options	**1** (appropriate)	**2**	**3**	**4**	**5** (not appropriate)

Once again thank you for your time and effort

Judgment Checklist – ACADEMIC EXTRACT VERSION

This checklist evaluates text/extract no:

What is required from you is your response to a set of statements using the Likert scale, the ratings of which are explained in each statement. You will need to have your test paper/text extract ready to complete the checklist. **Circle the number** that most closely expresses your point of view. Remember that there is no correct or incorrect answer. Please respond to each item.

A. Linguistic Demands

Overall text purpose

Since purpose and task both relate to the choice of text, a consideration of text type and topic is crucial to content validity.

1. The category that best describes the overall text purpose is…

	1	2	3	4	5
Options	Referential	Conative	Emotive	Poetic	Phatic
Examples	intended to inform	intended to persuade	intended to convey feelings or emotion	intended to entertain, delight, please	intended to keep in touch

Writer-reader relationship

The intended audience determines the extent of content knowledge that the writer can assume that the reader already has.

2. Identify the intended audience/reader of the text that is targeted by the writer.

	1	2
Options	Audience addressed	Audience invoked
Examples	is the intended reader	can be a fictitious reader, which often happens where the writer is writing for a rhetorical purpose

Discourse mode (Genre, rhetorical task, pattern of exposition, rhetorical organisation)

The reader's understanding of how texts are organized influences reading comprehension. In discourse mode, an argument or point of view or discussion is gradually built up.

Genre:

3. **Identify the most appropriate category for the text.**

	1	2	3	4
Options	text book	magazine/newspaper article	research/academic journal article	report

Rhetorical task:

4. **Identify the most appropriate category for the text.**

	1	2	3
Options	exposition	argumentation/persuasion/ evaluation	historical biographical/ autobiographical narrative

Pattern of exposition:

5. **Identify the pattern(s) used in the text *(You may decide that more than one option applies).***

	1	2	3	4	5	6	7	8	9
Options	define	describe	elaborate	illustrate	compare / contrast	classify	cause/ effect	problem /solution	justify

Rhetorical organisation:

6. **The organisational structure of the text is…**

	1	2	3	4	5
Options	Explicit				Not explicit

Functional resources

The illocutionary force of what is said. Examples include situations where a speaker has to persuade, advise, describe, etc

7. Identify the most appropriate category for the text.

Options	**1** Ideational	**2** Manipulative	**3** Heuristic	**4** Imaginative
Examples	Descriptions, classifications, explanations, and expressions of sorrow or anger	Requests, suggestions, commands, and warnings; rules, regulations and laws; greetings and leave-takings, compliments, insults, and apologies	for teaching and learning, problem solving, retention of information	Jokes, and use of figurative language and poetry

Grammatical resources (grammar, cohesion)

Grammatical forms are important for inferring the exact meaning intended by the writer.

Grammar:

8. The sentences in the text are:

Options	**1** mainly simple sentences	**2** a balance of simple and compound sentences	**3** mostly compound sentences	**4** a balance of compound and complex sentences	**5** mostly complex sentences

Cohesion

9. Throughout the text, are relations between the ideas explicitly marked through reference, conjunctions and connectors or are such relations not explicit?

Options	**1** (explicit)	**2**	**3**	**4**	**5** (not explicit)

Nature of information (Text abstractness)

The degree to which a particular text is posed in predominantly abstract or concrete terms.

10. Is the text concrete or abstract?

Options	1 (concrete)	2	3	4	5 (abstract)
Examples	The degrees of abstraction involved in the word 'chair' could be 'seat', then, 'furniture', then 'furnishings' and finally 'entity'.				

Content knowledge (General, cultural, language, religion)

What influence the test taker's background knowledge may have on the relative difficulty of a specific test task.

11. Is the topic of the text of general interest or does it require subject specific knowledge on the part of the reader?

Options	1 (general	2	3	4	5 (specific)

Cultural background:

12. Is the topic of the text culture-neutral or is it loaded with specific cultural content?

Options	1 (Culture neutral)	2	3	4	5 (cultural specific)

Language background:

13. The text is significantly easier to understand for readers from a specific first language background

Options	1 Strongly agree	2	3	4	5 Strongly disagree

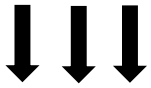
Religion knowledge:

14. Is the topic of the text religion-neutral or is it loaded with specific religious content?

Options	1 (religion neutral)	2	3	4	5 (religion specific)

Thank you for taking time to complete

(A. Linguistic Demands)



Please turn this page over and continue with questions on (B. Task Setting)

B. Task Setting

Channel of presentation

Comprehension is aided by information presented in more than one form. For example, text accompanied by a diagram, a picture or chart aids working memory.

15. The channel for the target situation requirements of the students being tested is…

Options	**1** (appropriate)	**2**	**3**	**4**	**5** (not appropriate)

Text length

Of sufficient length and size to provide ample opportunities for test takers to demonstrate their reading and comprehension. The actual length will depend on the purpose of testing.

16. The text length for the target situation requirements of the students being tested is…

Options	**1** (appropriate)	**2**	**3**	**4**	**5** (not appropriate)

Once again thank you for your time and effort

Appendix 9 The preliminary letter

Dear [Name of English Language HoC],

I am requesting your participation and the participation of your students and teachers in a short questionnaire that will investigate attitudes to current English language assessment in the Foundation Programmes at Colleges of Technology. This survey, which forms an important part of my doctoral research, is aimed at exploring the views of students and faculty members in the Foundation Programme regarding assessments methods in English language, with a view to making recommendations for improvement.

All volunteers will be asked to complete a short questionnaire. Student's questionnaire will have Arabic translation available. The questionnaire should take about 20 minutes to complete. All information you provide will be accessed only by the researcher. You are also assured that all data collected will be used for research purposes only. Your responses will be used in an anonymous way and you are hereby assured that your responses will be treated with strictest confidentially.

If you have any questions or concerns, please contact me by e-mail at: anwar-amar@hotmail.com

Your participation will be highly appreciated.

Anwar Al-Ismaili
PhD researcher
School of Social and International Studies
University of Bradford
United Kingdom

Appendix 10 Consent form

Consent Form

The table below presents a brief outline of the aims and purposes of this questionnaire. Please read it carefully before you agree to participate.

Title of Research	Ensuring the Context Validity of English Reading Tests for Academic Purposes (EAP)
Name Researcher	Anwar Amur Salim Al-Ismaili
Address for correspondence	School of Social and International Studies University of Bradford, United Kingdom West Yorkshire, BD7 1DP
Telephone	0044(0)1274232323
E-mail	anwar-amar@hotmail.com
General nature of the research	Investigating the English language assessment process and your views of its effectiveness
What is the expected involvement of the participant? How long the involvement will take?	Participants are expected to take about 1 hour to take a reading test and answer the questionnaire. The questionnaire has two parts Part (1): Questions asking about background information about you. Part (2) consists of 34 statements examining the key issues of second language assessment in your college. Participants are kindly asked to answer all parts.

My consent

I am aware that the purpose of this questionnaire is to make improvements to the current English language assessment practice. I understand that my responses will be used in a generalised way and with strict confidentiality and anonymity. I also understand that I have the right to request that my responses be withdrawn from the survey if I change my mind within 2 weeks.

By signing this consent form, I [*Your full name (in* BLOCK CAPITALS*)*]:_____ confirm that I understand the purpose of the questionnaire and hereby give my consent to participate in accordance with the information provided above.

Signed:_____ Date(dd/mm/yy):

_____

Appendix 11 TABLE 4.4 Total variance explained

Component	Initial Eigenvalues		
	Total	% of Variance	Cumulative %
1	5.591	16.445	16.445
2	2.654	7.807	24.252
3	1.960	5.766	30.018
4	1.772	5.213	35.230
5	1.525	4.485	39.715
6	1.463	4.302	44.017
7	1.277	3.755	47.772
8	1.215	3.574	51.346
9	1.139	3.349	54.695
10	1.072	3.152	57.847
11	1.037	3.050	60.897
12	.962	2.829	63.726
13	.901	2.650	66.376
14	.869	2.555	68.932
15	.824	2.422	71.354
16	.786	2.312	73.665
17	.731	2.149	75.814
18	.725	2.132	77.946
19	.684	2.012	79.958
20	.641	1.886	81.844
21	.628	1.847	83.691
22	.578	1.700	85.392
23	.555	1.632	87.023
24	.534	1.571	88.595

25	.510	1.501	90.096
26	.460	1.353	91.449
27	.450	1.324	92.773
28	.435	1.279	94.052
29	.419	1.233	95.286
30	.357	1.049	96.334
31	.347	1.020	97.355
32	.336	.990	98.344
33	.293	.861	99.205
34	.270	.795	100.000

Note. It should be noted that the numbering of the variables (component) does not correspond to the numbering of variable in this study but will be named in a later output.

Appendix 12 Table 4.8 Variables with strong loading on Component 2

#	Variables	Where in test paper (# item/question)	Coefficient result from Component Matrix	Where in questionnaire (statement)?	Actual cognitive processes used during the experiment	
					Likely process/es used (Agreement Responses)	**Likely alternative process/es used (Disagreement Responses)**
1.	Expeditious Search 2	11	.630	**S (9):** I had to quickly search the whole passage to find the information required to answer question 11.(65.3% X 34.7%)	Majority of the test takers were employing scanning expeditiously in order to answer the question	▪ Probably by careful reading ▪ Possible explanation L1-L2 transfer word loaded /saliency of letters than English learners spend more time on letters and words than they do at global level

2.	Expeditious Search 3	12	.622	**S (10):** I had to quickly search the whole passage to find the information required to answer question 12.(57.5% X 42.5%)		
3.	Expeditious Search 4	13	.595	**S (11):** I had to quickly search the whole passage to find the information required to answer question 13.(62.9% X 37.1%)		
4.	Expeditious Search 1	10	.495	**S (8):** I had to quickly search the whole passage to find the information required to answer question 10. (71.8% X 28.2%)		

5.	Lexical Resources 3	Relates to test text Line 7	.482	**S (15):** I needed to read beyond the sentence in line 7 in order to understand the meaning of 'the climate is tropical and humid'. (62.3% x 37.7%)	Inferred the meaning of unfamiliar words as intended It supports what has been said in the previous section (Grammatical resource - syntactic) that it indicates that the students inappropriately applied the incorrect strategy which resulted in their attempts to read at a global level whereas the repeated use of the conjunction 'and' (syntactical resource),which here does not have the connotation of 'additional' but more like 'or', clearly implies that it was possible to infer the meaning at local level without going beyond the sentence. In fact, going beyond the sentence in this case provided no	▪ knew already the meaning of 'extensively' (used background knowledge) or else ▪ guessed the meaning in a multiple choice context

					additional aids to completing the task	
6.	Discernment1	5 & 6	.386	**S (2):** I found it difficult to decide whether to skim (fast read) or read carefully the whole passage in order to answer questions. (64% X 36%)	▪ Applied discernment but still found it difficult or else ▪ Applied discernment but to a lesser degree	▪ Probably more able to organize their mental frame or grouping relationships ▪ Read selectively according to goals
7.	Channel of Presentation 1	Relates to test text	.348	**S (24):** A time line of main events would have aided and supported my passage comprehension.(65.9% X 43.1%)	May indicate that reading the text alone was somewhat challenging and that additional aid to comprehension was desirable,,	Possibly able to rely lexis and grammatical resources for meaning as well as draw on background knowledge
8.	Rubric 1	Part D in test paper 10, 11, 12, 13	.315	**S (7):** I was unsure whether long detailed answers or short answers were required in answering Part D. (63.9% x 36.1%)	Rubrics were clear and aided the majority of students understanding of the task. One exception is statement 7 which did not indicate the level of detail required	▪ Might have used familiarity with practice test ▪ Guessing what was required

Appendix 13 Table 4.9 Variables with strong loading on Component 1

#	Variables	Where in test paper (Item/question?)	Coefficient result from Component Matrix	Where in questionnaire (Statement?)	Actual cognitive processes used during the experiment	
					Likely process/es used (Agreement Responses)	Likely alternative process/es used (Disagreement Responses)
1.	Discourse Mode 2	Relates to test text	.628	31: I think the way the paragraphs are arranged helped me to understand the passage. (77.3% X 22.7%)	▪ Test takers were generally able to use cohesive devices in order to grasp the overall meaning (chronological structures of paragraphs) ▪ Also, may have used the available genre of historical events to comprehend the task	▪ Probably did not use available cohesive devices effectively ▪ Probably may have drawn on their background knowledge
2.	Rubric 2	Relates to test text	.625	19:I found the test instructions easy to understand. (78.3 x 21.7%)	Rubrics were clear and aided the majority of students understanding of the task. One exception is statement 7 which did not indicate the level of detail required	▪ Might have used familiarity with practice test ▪ Guessing what was required

3.	Discourse Mode 1	Relates to test text	.613	29: I think the way the paragraphs are arranged helped me to understand the passage. (77.3% X 22.7%)	Same as above in 1	Same as above in 1
4.	Content Knowledge	Relates to test text	.607	33: The article was mostly familiar to me due to my knowledge of the history of the connections between Zanzibar and Oman. (75.2% X 24.8 %)	Same as above in 1	Same as above in 1

5.	Response Method 2	The use of varieties of question type but not specific to one type	.592	21: I found the variety of question types (e.g. true/false, short answer, gap filling, etc.) helpful in allowing me to show my skills. (80.7% X 19.3%)	May have read carefully at global level instead of searching expeditiously to answer a deletion gap fill For statement 21: May have drawn on their many skills to respond to different types of response methods including self-management and prioritizing.	▪ May have read expeditiously to answer a deletion gap filling exercise ▪ For statement 21: May have not applied appropriate self-management skills
6.	Overall Passage Purpose	Relates to test text	.571	28: Recognising the overall purpose of the passage helped my understanding. (70.7% X 29.3%)	▪ Employed the strategy of overall passage purpose to comprehend the passage ▪ Probably by expeditious reading they were able to identify the overall purpose which aided comprehension	▪ It is likely that they would have disagreed on the basis that they have identified the purpose of the passage and found this to be helpful. ▪ The more likely explanation is the misuse of careful reading of the entire passage which rendered it difficult to identify the purpose ▪ Or they did use expeditious reading but did not do so appropriately

7.	Nature of Information	Relates to test text	.553	32: I found most of the words on the passages were factual e.g. car, apple.(74.2% X 23.8%)	▪ Most likely used expeditious reading appropriately to know and understand the meaning ▪ Also, likely used strategies such as prediction and hypothesizing to understand some of the abstract words ▪ Perhaps also drawn on their cultural background	▪ Probably used careful reading most of the time ▪ Perhaps have not yet developed strategies of prediction or hypothesizing ▪ Use of cultural background with lesser degree
8.	Grammatical Resources 2	Relates to test text 'it' in line 7	.542	17: I could easily understand 'it' referred to in line 7. (74% X 26%)	▪ (for 17 & 18) Probably were reading carefully at global level which suggests that they used available grammar resources appropriately. ▪ Also it means that they were connecting ideas in different sentences and relating them to each other	▪ Possibly they did an expeditious reading at global level instead of careful reading ▪ Possibly stuck focusing at lexis level
9.	Writer-Reader	Relates to test text	.535	30: This passage is suitable for Omani students.(76.2% X 23.8%)	Test takers were probably able to draw on their prior knowledge/content knowledge	Probably were not able to connect the passage to their existing schemata and prior knowledge

10.	Channel of Presentation 2	Relates to test text Use of headings	.520	34: I was easily able to understand the passage even though there were no headings. (67.3% X 32.7%)	Reliance on the text alone was adequate for comprehension, possibly aided by prior knowledge or test-wiseness (elimination process)	▪ Used prior knowledge with lesser degree than those who agreed Not fully able to use clues in the passage due to probably lack of awareness of syntax or shortage of lexical resources
11.	Scanning Expeditiously	1, 2, 3, 4	.519	1: I was quickly able to find the information required to answer the questions in PART A.	Majority of the test takers were employing scanning expeditiously in order to answer the question	▪ Probably by careful reading ▪ Possible explanation L1-L2 transfer word loaded /saliency of letters than English learners spend more time on letters and words than they do at global level
12.	Weighting	Relates to all questions	.504	20: I think the number of marks given for each question was appropriate. (77.2% X 22.8%)	Test takers may have used discernment to prioritize or self-management for time allocation to a task	It is likely that disagreement indicates underdeveloped skills of discernment and time management by considering weighting and knowledge of criteria.

13.	Careful Local	9	.499	5: I was able to answer question 9 by carefully reading the sentences in line 17 and 18. (66.3% x33.7%)	Appropriately applied careful reading at the local level	▪ May have applied careful reading at local level but did so less effectively or else ▪ Applied Careful reading at global instead of local level
14.	Grammatical Resources 3	Could be 6 But 6 came low less than .4 under Grammatical resource 1	.477	18: I found it easy to decide which of the three options best express the main idea of Paragraph 1 because the sentences were short. (62% X 38%)	▪ (for 17 & 18) Probably were reading carefully at global level which suggests that they used available grammar resources appropriately. Also it means that they were connecting ideas in different sentences and relating them to each other	▪ Possibly they did an expeditious reading at global level instead of careful reading ▪ Possibly stuck focusing at lexis level

15.	Knowledge of Criteria	Relates to all questions	.457	22: It was helpful to know how many marks were allocated to each item. (76.8%, X 23.2%)	Test takers may have used discernment to prioritize or self-management for time allocation to a task	It is likely that disagreement indicates underdeveloped skills of discernment and time management by considering weighting and knowledge of criteria.
16.	Passage Length	Relates to the test text	.437	25: I did not need to read every word of the passage in order to understand its meaning. (61.8% X 38.2%)	Test takers most likely used expeditious reading (skimming/search/scanning) to cope with and understand a longer text	▪ Test takers probably used using careful reading inappropriately. Not using expeditious reading appropriately

17.	Careful Global	8 This item also examines lexical resource 1 which came less than .4	.407	4: I had to read other sentences carefully in addition to the sentence on line 16 in order to answer question 8. (65.3% x 34.7%)	Appropriately applied careful reading at the local level This may apply to statement 5 not 4…just check	▪ May have applied careful reading at local level but did so less effectively or else ▪ Applied Careful reading at global instead of local level
18.	Time Constraint	Relates to all questions	.400	27: 40 minutes was sufficient time to answer all the questions. (62.8% & 37.2%)	Test takers were able to discern whether expeditious or careful reading was required	▪ Probably used careful reading almost all the time and did not use expeditious reading appropriately in terms of time management, and ▪ may consequently may have found time management difficult

19.	Discernment 2	Relates to the test text and all questions	.372	26: I looked at the questions first before deciding whether to read the passage carefully or quickly. (73.2% X 26.8%)	▪ Applied discernment but still found it difficult or else ▪ Applied discernment but to a lesser degree	▪ Probably more able to organize their mental frame or grouping relationships ▪ Read selectively according to goals

Appendix 14

Table 4.14 Questionnaire statement 1 and test item 1

Test Statistics[a]

	Questionnaire Score Statement 1 Component 1
Mann-Whitney U	1422.000
Z	-.339
Asymp. Sig. (2-tailed)	.735

a. Grouping Variable: Item 1 Scores for statement 1

Table 4.15 Questionnaire statement 1 and test item 2

Test Statistics[a]

	Questionnaire Score Statement 1 Component 1
Mann-Whitney U	2592.000
Z	-.486
Asymp. Sig. (2-tailed)	.627

a. Grouping Variable: Item 2 Scores for statement 1

Table 4.16 Questionnaire Statement 1 and test item 3

Test Statistics[a]

	Questionnaire Score Statement 1 Component 1
Mann-Whitney U	2521.000
Z	-1.906
Asymp. Sig. (2-tailed)	.057

a. Grouping Variable: Item 3 Scores for statement 1

Table 4.17 Questionnaire Statement 1 and test item 4

Test Statistics[a]

	Questionnaire Score Statement 1 Component 1
Mann-Whitney U	4454.000
Z	-1.782
Asymp. Sig. (2-tailed)	.075

a. Grouping Variable: Item 4 Scores for statement 1